# Auditory Kernels for Representing Degraded Speech

## Auditory Kernels in an Efficient Representation of Degraded Speech

**Baturalp Karslioglu**[1]

**Supervisor(s): Jorge Martinez**[1]**, Dimme de Groot**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Baturalp Karslioglu

Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez, Dimme de Groot, David Tax

## Abstract

We explore the use of biologically inspired auditory kernels—learned from sparse coding on (clean) read speech—to analyze and reconstruct signals degraded with additive noise. Auditory kernels mimic spectrotemporal filters in the human auditory system, offering insight into how structured acoustic signals can be internally represented and selectively preserved. Our study applies an auditory kernel-based matching pursuit reconstruction framework to clean, degraded, and standalone noise audio, investigating kernel activation patterns across input types. The findings reveal kernel selectivity; structured signals like speech activate a common subset of kernels, while unstructured noise elicits distinct, less overlapping activations, allowing for more effective separation and implicit denoising. This selectivity results in implicit denoising, preserving intelligibility and perceptual quality even under degradation. By quantifying this behavior across noise types and SNR levels, we show that auditory kernels not only support robust signal reconstruction but also offer a biologically grounded, explainable mechanism for speech enhancement. These insights advance the use of sparse auditory models in both neuroscience and signal processing, motivating future work on adaptive or context-aware dictionaries.


## 1 Introduction

Auditory signals are the complicated sound waves we perceive and interpret through our hearing system. Internally, the auditory system transforms these signals into sparse, structured representations—capturing the most informative aspects of sound while discarding redundancy [1]. Understanding how these internal auditory representations work is a central question in both neuroscience and audio signal processing. One promising framework to explore this is through **auditory kernels**—basis functions or filters that resemble how biological systems, like the human cochlea and cortex, respond to sound.

A particularly influential approach has been to learn sparse dictionaries that emulate properties of early auditory processing. Michael Lewicki [1] provided a key foundation by showing that many cochlear filter properties can emerge from the principle of efficient coding of natural sounds. Later, Smith and Lewicki showed that a good representation of human speech can be represented as how humans hear. Using those gammatone filters that replicate the cochlea, auditory kernels are constructed. Moreover, it is shown that these kernels can be used efficiently to encode and decode human speech. [2]

Building on this foundation, recent studies, for example, Ming and Holt [3] have extended these ideas to investigate how such kernels behave in the context of degraded inputs or alternative sound sources. Other neuro-scientific studies [4; 5] supports that cortical responses remain selective for speech-like acoustic structures, even under heavy noise, supporting the idea that the auditory system inherently prioritizes structured inputs.

However, it remains unclear how these learned representations respond to purely unstructured input, such as additive white noise, or specific degradation patterns applied to speech. This gap limits our understanding of the generalizability and selectivity of these auditory kernels.

This paper addresses the following research questions:

- **RQ 1** *How much does the auditory kernel reconstruction selectively reconstruct speech-like patterns even in conditions where speech is degraded?*
  - *What are the signs of implicit denoising—i.e., removing non-speech structure through a sparse kernel matching?*
  - *How well does the auditory kernel reconstruction work under different noise types in terms of signal-to-residual ratio (SRR)?*
- **RQ 2** *How does the quality of reconstructed signals evolve across different noise types and signal-to-noise ratio (SNR) in terms of perceptible quality and intelligibility?*
  - *What are the results of comparing clean and degraded reconstructed speeches across different noise types and SNRs in terms of quality (perceptibility) and intelligibility metric scores?*
- **RQ 3** *How can we quantify the selectivity of auditory kernel activations concerning different noise types and degraded speeches with those noises?*
  - *What different kernels are activated for speech versus noise, and what are the similar patterns across different noise types?*

Main contribution of this paper is to analyze different categories of auditory input using pre-trained (on clean speech) auditory kernels. We systematically evaluate kernel activations and reconstruction output, providing insight into the internal structure captured by these dictionaries. Our findings contribute to the understanding of kernel specialization and raise new directions for analyzing auditory representations in neural and artificial systems.

The rest of the paper is structured as follows. Section 2 discusses related work in auditory modeling and sparse coding. Section 3 presents our methodology, including degradation strategies and evaluation metrics. Section 4 explains the potential contribution of this research. Section 5 gives more details on the experiment and how it was conducted. Section 6 shows the results of the experiment, and Section 7 talks about Responsible Research. Section 8 is the discussion part. Section 9 concludes with implications and future research directions.

## 2 Related Work

Sparse coding has emerged as a biologically inspired approach for understanding how sensory systems may efficiently represent complex stimuli. In the auditory domain, a foundational contribution by Smith and Lewicki (2006) demonstrated that sparse overcomplete dictionaries learned

from natural sound statistics yield time-frequency kernels resembling spectrotemporal receptive fields observed in the mammalian auditory cortex [2]. Their work provided a strong theoretical basis for modeling auditory representations using unsupervised learning, revealing how auditory systems might be tuned to encode structure present in natural environments.

Lewicki (2002) provided earlier foundational insight by demonstrating that many cochlear filter properties can emerge from the principle of efficient coding of natural sounds [1]. This work solidified the theoretical foundation for applying sparse coding to auditory perception, showing that sparse representations align closely with how the auditory system reduces redundancy while preserving perceptual structure.

Carlson et al. (2012) further extended this line of research by learning overcomplete sparse codes for speech that closely resembled the receptive fields of mid-level auditory neurons [6]. Their findings support the notion that sparse auditory kernels are not just cochlear approximations but also capture cortical-level processing dynamics. Similarly, Chi et al. (2005) proposed a multiresolution model that mimics both cochlear and cortical processing using gammatone and modulation filterbanks [7], offering a biologically grounded framework that aligns with psychoacoustic and neurophysiological observations.

Lewicki (2010) further expanded this view by emphasizing the relevance of signal structure and redundancy reduction, arguing that the auditory system likely performs an efficient decomposition of sound into sparse, informative features [8]. These biologically motivated models contrast with purely engineering-based representations by mimicking perceptual priors embedded in the human auditory system. Lewicki's work supports the Efficient Coding Theorem, which proposes that sensory systems are adapted to efficiently represent natural stimuli by reducing redundancy [2]. In the auditory system, this means using sparse, speech-like features that reflect how humans naturally produce and perceive sound.

Building on this foundation, Ming and Holt (2009) examined sparse auditory representations under noisy listening conditions and showed that listeners may rely on sparse temporal cues for perceptual judgments, even when traditional spectral fidelity is compromised [3]. Means, even in noisy environments, people can still make sense of speech by picking up on just a few strong timing-based cues, thanks to the way sparse representations highlight the most perceptually relevant parts of sound.

From a neuroscience perspective, Mesgarani et al. (2014) and Souffi et al. (2021) provide evidence that cortical and distributed auditory responses remain selective to speech-like patterns even in noisy environments. Their results suggest that the auditory system preserves structured, behaviorally relevant features of speech despite additive or reverberant distortions [4; 5]. Complementing this, Miettinen et al. (2010) showed that the human auditory cortex is particularly responsive to harmonic and periodic components in degraded signals, whether speech or non-speech, highlighting the general importance of structured spectral patterns in auditory processing [9].

Additionally, Yao's study on speech perception under aviation noise environments revealed how speech intelligibility is disproportionately affected by different noise types, underscoring the need for signal representations that adapt to real-world degradations [10]. While Sigg's study demonstrated that sparse dictionaries can explicitly improve speech quality by suppressing noise through learned structure [11]. Together, these studies motivate our kernel-based approach to examine how auditory dictionaries behave under diverse degradations.

This study builds on this body of work by empirically examining how auditory kernels, learned from natural sound statistics, behave under additive noise when used for speech reconstruction. By using biologically informed kernels (dictionaries) derived from natural auditory scenes, we aim to understand better the robustness and perceptual relevance embedded in these representations, rather than proposing a new algorithm.

## 3 Methodology and Research Design

This section outlines the methodology and experimental design used to evaluate the effectiveness of auditory kernels in reconstructing speech signals under additive noise conditions. Our research aims to understand how auditory kernels perform in preserving intelligibility and quality of speech in noisy environments. We also investigate whether these kernels offer any denoising capacity or display patterns that reveal sensitivity to specific noise types.

### 3.1 Application and Objective

For this study, we use a **train station environment** as the representative environment, due to its rich and diverse background noise profile. This includes noise types that are modeled as additive noise types, such as overlapping human speech (babble), mechanical noise from trains, periodic station announcements, and other ambient station sounds. The goal of this research is to analyze how effectively auditory kernels reconstruct degraded speech in those environments and also to analyze the patterns across different environments—noise types. Measures for efficiency for this research are mainly how intelligible and high-quality speech remains when these noises are present.

### 3.2 Dataset and Data Preparation

To simulate realistic scenarios encountered in a train station, we selected clean speech signals from the Microsoft Scalable Noisy Dataset (MSND) [12]. Specifically, we used recordings from 25 male and 25 female speakers, with two utterances per speaker, resulting in 100 speech samples in total. All signals were sampled at 16 kHz for consistency.

The additive noise types introduced to clean speech are as follows:

- **Babble Noise:** A mixture of human voices from MSND that usually occur in train stations.
- **Station Announcement:** Pre-recorded announcements that usually occur in train stations from the MSND.
- **Train Arrival Noise:** Noise of train coming and opening doors, recorded at Delft Station, South Holland.

- **White Noise:** Synthetic white Gaussian noise generated for baseline comparison.

All speech samples are **Root-mean square (RMS) normalized** before degradation to ensure consistent amplitude levels. The RMS normalization is computed using [13]:

$$\text{RMS}(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{1}$$

Additive noise is applied at four Signal-to-Noise Ratio (SNR) levels: -5 dB, 0 dB, 5 dB, and 10 dB. Give a variety of ranges and keep the speech hearable. SNR is calculated using:

$$\text{SNR} = 10 \cdot \log_{10}\left(\frac{||s||_2^2}{||n||_2^2}\right) \tag{2}$$

where $s$ is the clean signal and $n$ is the noise signal. The notation $\|\cdot\|_2$ denotes the **Euclidean norm**, which computes the square root of the sum of squared values [14].

### 3.3 Matching Pursuit and Kernel-Based Reconstruction

Matching Pursuit (MP) is a greedy, iterative algorithm introduced by Mallat and Zhang [15] for sparse signal representation using time-frequency dictionaries. It aims to approximate a signal by selecting a series of time-shifted and scaled kernels (or atoms) from a predefined dictionary. The process is particularly suitable for auditory modeling, as it captures the key temporal-spectral structures in speech through a compact representation.

Let $x(t)$ be the degraded speech signal. MP approximates it as:

$$x(t) \approx \sum_{i=1}^{K} a_i g_{\gamma_i}(t) \tag{3}$$

where each $g_{\gamma_i}(t)$ is an kernel from the dictionary $\mathcal{D}$, defined by a kernel function $g$ with parameters $\gamma_i = \{\tau_i, \phi_i, s_i\}$ indicating time shift, frequency, and scale. The coefficient $a_i$ represents the energy projection onto the selected kernel. [2]

The algorithm proceeds as follows:

1. Initialize the residual $R^0 = x(t)$.
2. At each iteration $k$, select the kernel $g_{\gamma_k}$ that maximizes the inner product with the residual:

$$g_{\gamma_k} = \arg\max_{g \in \mathcal{D}} |\langle R^{k-1}, g\rangle| \tag{4}$$

3. Update the residual:

$$R^k = R^{k-1} - \langle R^{k-1}, g_{\gamma_k}\rangle g_{\gamma_k} \tag{5}$$

4. Repeat until a stopping criterion is met (e.g., a maximum number of kernels $K$ or the amplitude of the selected coefficient $|\langle R^k, g_{\gamma_k}\rangle|$ is lower than a set minimum).

In this research implementation, the dictionary is composed of **auditory kernels** (`kernels_15040.jld2` are kernels that are trained on clean speech by our supervisor [16])—time-frequency localized basis functions learned from natural speech data, inspired by early auditory processing models [2], [3]. These kernels often resemble modulated sinusoids or frequency sweeps, matching the spectrotemporal features observed in cochlear and cortical responses, and offer a meaningful decomposition of speech [8]. Furthermore, Matching Pursuit (MP) is chosen as the reconstruction method not only for its interpretability but also because exact sparse coding is known to be NP-hard [17], making MP a practical approximation method.

The reconstruction is achieved by summing the selected kernels. Because of the variability in speech and noise characteristics, the number of kernels $K$ used for each signal may vary. However, this sparse representation enables us to analyze how the auditory system prioritizes certain sound features under noise and whether specific patterns emerge when dealing with different types of additive noise.

The selected kernel parameters (e.g., type, scale, position) are stored in order to use them in our subsequent evaluation and pattern analysis to study how noise influences kernel selection.

### 3.4 Evaluation Metrics

After reconstructing the speech signals, metrics are necessary to evaluate the reconstructed signal. To evaluate reconstruction quality, we focus on both **intelligibility** and the quality of the sound (**perceptibility**). **Intelligibility** refers to how well the linguistic content of speech can be understood by a listener, regardless of audio quality. It is closely related to the preservation of phonetic and temporal features necessary for comprehension. On the other hand, **perceptibility** (or perceptual quality) focuses on the overall auditory experience of the signal — how natural, pleasant, or distorted it sounds, even if the words are understandable. These two dimensions are complementary and essential for assessing the effectiveness of reconstruction methods in real-world noisy environments.

We use the following three objective metrics, each designed to capture one of these perceptual aspects:

- **PESQ (Perceptual Evaluation of Speech Quality):** PESQ is developed to objectively assess the perceived quality of speech, especially in telephony. It compares a reference (clean) signal and a degraded signal by modeling perceptual transformations in the human auditory system. The core computation involves mapping both the reference and degraded signal into an internal representation, followed by a disturbance processing stage that captures audible differences [18], [19]. In our implementation, a Python library is used to evaluate the final MOS-LQO (Mean Opinion Score – Listening Quality Objective) score from these disturbances [20].
- **ViSQOL (Virtual Speech Quality Objective Listener):** ViSQOL is a full-reference, signal-based metric designed to predict the perceived quality of speech by modeling human auditory perception mechanisms [21]. It operates by converting both the clean (reference) and degraded signals into spectro-temporal representations using the short-time Fourier transform (STFT), which are then divided into patches. The similarity between corresponding patches is measured using a similarity index measure (Neurogram Similarity Index Measure).

The average similarity score across all patches is then mapped to a Mean Opinion Score (MOS) scale ranging from 1 (bad) to 5 (excellent):

- **STOI (Short-Time Objective Intelligibility):** Estimates speech intelligibility by computing the correlation between time-aligned short-time spectral representations of clean and degraded speech signals [22]. Both signals are first decomposed into overlapping short-time frames and passed through a 1/3-octave band filter. The average correlation over all frames forms the final STOI score.

  A STOI score ranges from 0 (completely unintelligible) to 1 (perfect intelligibility), and higher scores indicate better preservation of intelligibility in noisy environments.

These metrics were selected based on their standardization, practical relevance, and complementarity. While ViSQOL is included for completeness, our main focus is on **PESQ** and **STOI**, as they directly align with our evaluation goals and are used in our reported results. **PESQ**, used to be standardized as ITU-T P.862, is widely used in telecom and codec evaluation to assess perceptual quality [23], while **STOI** is a standard intelligibility metric in scientific studies and shows strong correlation with human understanding. Together, they provide a balanced assessment of how natural the reconstructed speech sounds and how understandable it remains under degradation and reconstruction with kernels.

Since all these metrics need a reference signal and a degraded signal (signal to compare), metrics under three pairwise comparisons are computed as:

1. Clean speech vs. Reconstructed Clean speech with kernels: to measure the baseline performance under these metrics.

2. Degraded vs. Reconstructed Degraded speech with kernels: to measure the performance change when the speech is degraded.

3. Reconstructed Clean speech with kernels vs. Reconstructed Degraded speech with kernels: to isolate the effect of degradation on the reconstruction quality. This comparison also helps assess whether the reconstruction process with auditory kernels introduces any denoising effect—i.e., if the reconstructed degraded speech is perceptually or intelligibly closer to the clean version than the original degraded input. The reason is mainly to support Research Question 2 and contribute to Research Question 1.

Comparisons do not include Clean vs. Reconstructed Degraded because reconstruction inevitably introduces some loss, even for clean signals. Comparing Reconstructed Clean with Reconstructed Degraded ensures a more balanced evaluation, as both undergo the same processing. This allows us to better isolate the impact of degradation and assess whether auditory kernel-based reconstruction introduces any denoising effect.

# 4 Behavioral Analysis of Auditory Kernel-Based Matching Pursuit under Noise

This section synthesizes the theoretical and experimental insights gained from applying Matching Pursuit with auditory kernels for reconstructing speech under additive noise. Rather than proposing a new algorithm, our contribution lies in analyzing how an existing signal decomposition method—Matching Pursuit—behaves under additive noise when paired with biologically inspired auditory kernels, and what this behavior reveals about perceptual and intelligibility-oriented robustness in speech reconstruction.

## 4.1 Experimental Insights on Reconstruction Robustness

The application of PESQ and STOI across various noise conditions provides insight into how auditory kernel-based Matching Pursuit performs from both perceptual and intelligibility standpoints. While quantitative results are discussed in the next section, the following conceptual trends can be inferred from our experimental framework:

- **Perceptual Quality is Partially Retained:** The use of sparse auditory kernels appears capable of preserving mid-level acoustic structures, particularly in environments with structured noise (e.g., announcements). This suggests that biologically inspired representations may inherently encode some robustness to background interference, making them promising for real-world applications.
- **Intelligibility Degrades More Gradually:** Despite severe input degradation, intelligibility scores (STOI) decline more gradually compared to perceptual quality (PESQ). This suggests that the auditory kernel representation prioritizes temporally and linguistically important cues, making it resilient in noisy environments. Such properties could inform future speech compression or enhancement methods that emphasize intelligibility over raw audio fidelity.

These insights, while preliminary, suggest that auditory kernel-based representations offer a promising direction for speech analysis and reconstruction under challenging acoustic conditions. Future work could more directly model or enhance these denoising and robustness properties through targeted dictionary learning or adaptive kernel selection.

## 4.2 Theoretical Implications of Kernel Selection Behavior

Beyond performance metrics, auditory kernel speech reconstruction-due to the patterns kernels selects-provides deeper insights into the nature of auditory signal processing. Some expected theoretical patterns include:

- **Noise-Selective Representation:** Different noise types (e.g., stationary vs non-stationary) are likely to interfere with specific spectral-temporal patterns. As a result, the types of kernels selected during reconstruction

may shift, especially in frequency and duration, potentially highlighting how noise suppresses or alters salient features.

- **Soft Denoising through Sparse Matching:** Because Matching Pursuit selects kernels based on best-matching segments, it may naturally ignore noise-like structures that are not well-represented in the dictionary [15]. This results in a form of selective encoding that prefers structured, speech-like patterns—echoing how biological auditory systems suppress background noise through attention and adaptation mechanisms [2].

These patterns, while theoretical at this stage, suggest that auditory kernel dictionaries embed useful priors for robust speech representation, especially in challenging acoustic conditions.

# 5 Experimental Setup

This section outlines the evaluation protocol for testing how auditory kernel-based reconstruction performs under realistic additive noise conditions, mimicking a train station environment.

## 5.1 Dataset and Degradation Conditions

All clean speech signals were selected from the **MS-SNSD**[12], using a subset of recordings sampled at 16 kHz and normalized in amplitude. We applied four types of additive noise to generate degraded signals at various **signal-to-noise ratios (SNRs)** as described earlier.

These noise types were selected to *span a range of spectral and temporal complexities*, mimicking realistic auditory scenes. For example, **babble noise** reflects overlapping human speech, while **announcement recordings** may exhibit a speech-like structure that competes with target utterances. **Incoming train** sounds introduce broadband dynamic interference and high amplitude, while the **white noise** is just a stationary noise.

## 5.2 Kernel Dictionary and Reconstruction Procedure

We use a single, biologically inspired kernel dictionary composed of **32 learned auditory kernels**, trained on clean speech from the TIMIT corpus using sparse coding. These kernels are designed to capture common spectrotemporal patterns in speech and extend traditional auditory models by learning from data.

The design builds on the intuition behind **gammatone filters**—which model cochlear frequency selectivity [24]—but instead of using fixed filters, the kernels are learned to reflect recurring structure in (clean) read speech better.

Reconstruction is performed using the Matching Pursuit algorithm, which iteratively selects the kernel that best matches the residual signal. This continues until the **maximum inner product** between the residual and any kernel falls below **0.1**, ensuring that reconstruction proceeds only while meaningful structure remains in the signal. This approach yields sparse, interpretable reconstructions with consistent stopping conditions across different signals.

This same dictionary is used to reconstruct **clean speech**, **degraded speech**, and **noise-only inputs**—enabling analysis of kernel selectivity across conditions in later sections—for robustness and consistency.

## 5.3 Evaluation Metrics

To quantify reconstruction quality across speech quality (perceptibility) and intelligibility axes, we used two out of three standard objective metrics discussed:

- **PESQ** (Perceptual Evaluation of Speech Quality): for narrowband speech quality approximation [18]
- **STOI** (Short-Time Objective Intelligibility): for intelligibility estimation based on temporal envelope correlations [22]

Each metric was applied across multiple pairings as discussed in Section 3.

## 5.4 Experimental Objectives

The experimental procedure was structured to answer the subquestions of the Research Questions as discussed in Section 1:

- RQ 1: *What are the signs of implicit denoising* and, *How well does the auditory kernel reconstruction work under different noise types in terms of signal-to-residual ratio (SRR)?*
- RQ 2: *What are the results of comparing clean and degraded reconstructed speeches across different noise types and SNRs in terms of PESQ and STOI scores?*
- RQ 3: *What different kernels are activated for speech versus noise, and what are the similar patterns across different noise types?*

# 6 Results

This section presents the experimental findings of applying auditory kernel-based reconstruction to speech signals degraded with various types of noise under different SNR conditions. The results are organized around three key analysis perspectives: perceptual/intelligibility quality metrics, reconstruction process comparison along 4 different noise types, and activation of kernels compared. All results presented here are computed by averaging across 100 speech samples per condition from the dataset, as described in Section 5, ensuring robustness and generality of the observed trends.

## 6.1 Average Intelligibility and Perceptibility Scores

Results shows the average PESQ and STOI scores under two SNR conditions (10 dB and 0 dB), illustrating how perceptual quality and intelligibility is preserved or degraded depending on the noise level and type. The remaining plots can be found in the Appendix A. These results support the objective of the **Research Question 2**.

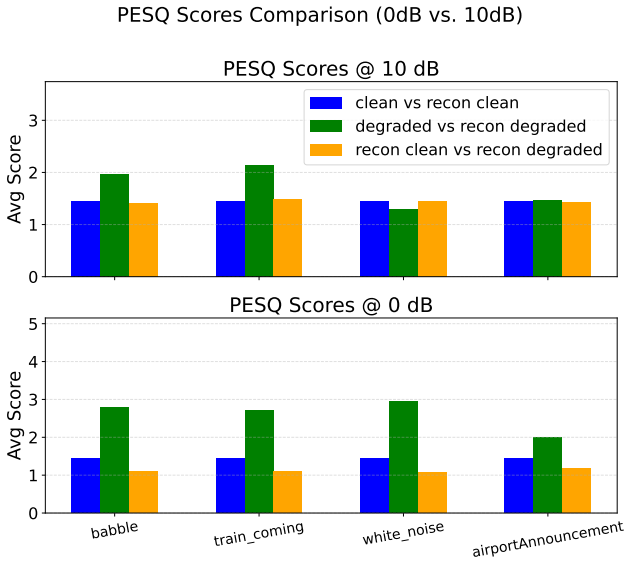



Figure 1: Averaged PESQ scores across four noise types under 10 dB and 0 dB SNR conditions.

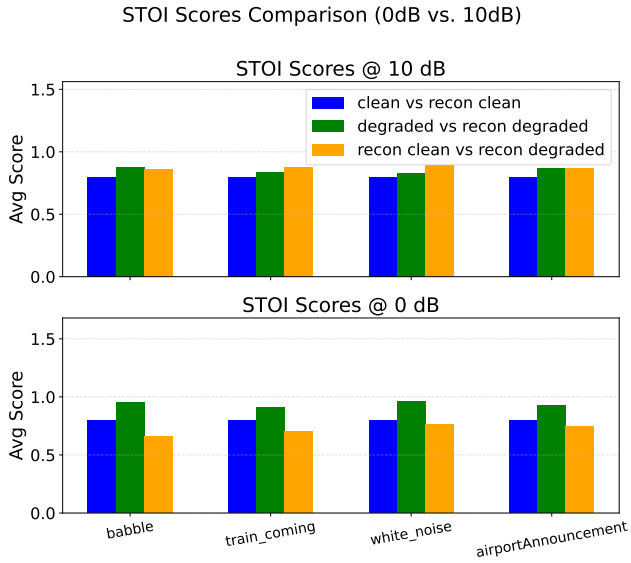



Figure 2: Averaged STOI scores across four noise types under 10 dB and 0 dB SNR conditions.

**PESQ Analysis:** Figure 1 PESQ scores shows that kernel reconstruction preserve the quality of the signal in most cases and give a hint of a denoising application.

At **high SNRs**, reconstructed degraded signals sometimes outperform the clean baseline in PESQ (yellow bars higher than blue). This is especially evident in train_coming and white_noise conditions. The result implies that auditory kernels may reduce perceptual noise artifacts, effectively applying *light denoising* even without explicit modeling.

At **low SNRs**, PESQ scores drop for yellow bars and increase for green bars, shows that noise is captured more as expected in lower SNRs. However, reconstructed signals still maintain a perceptually meaningful quality level (typically above 1.0) and are still close to the baseline (blue bar).

**STOI Analysis:** Figure 2 STOI scores show that across all SNRs, auditory kernel reconstruction preserves intelligibility well. Scores for reconstructed degraded are consistently high (around 0.9), showing the reconstruction process preserves intelligibility well under moderate noise conditions.

At **higher SNRs**, intelligibility of reconstructed degraded speech remains high. The "recon clean vs. recon degraded" bars (yellow) are often higher than the clean baseline (blue), indicating that auditory kernel reconstruction introduces a *denoising* on low-energy noises.

At **lower SNRs**, the yellow bars drop more noticeably, but the overall difference remains moderate. This suggests that even in adverse noise conditions, the auditory kernels preserve core speech patterns.

We also observe that **white_noise** and **train_coming** conditions yield better STOI scores when compared with the clean speech at overall in respect to **babble** and **airportAnnouncement**. This likely reflects the fact that babble and announcements are speech-like, making it harder for the kernel dictionary (trained on clean speech) to distinguish between target and noise.

## 6.2 Reconstruction Process Under Noise Types

To evaluate how much of the original signal energy is captured during reconstruction, we use the Signal-to-Residual Ratio (SRR) as a frame-wise metric. We fix the kernel budget to 500 kernels per second to assess reconstruction efficiency under different noise types and SNR conditions.

SRR is computed as:

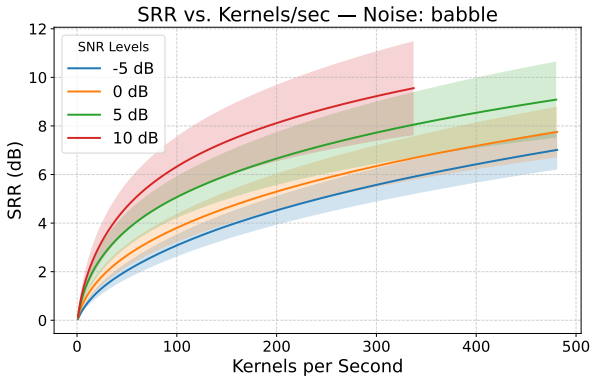
$$\mathrm{SRR}(k) = 20 \cdot \log_{10}\left(\frac{\|y\|_2}{\|r_k\|_2}\right) \tag{6}$$

where $y$ is the original input signal and $r_k$ is the residual signal after reconstructing with the first $k$ kernels. The notation $\|\cdot\|_2$ denotes the **Euclidean norm**, which computes the square root of the sum of squared values.

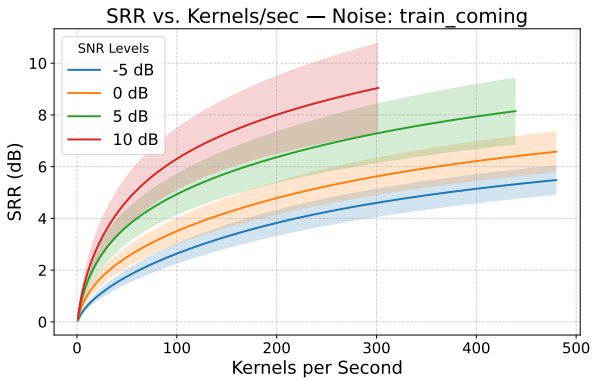
This metric quantifies how much of the original signal energy is captured as more kernels are used. A higher SRR indicates a more faithful reconstruction. Figure 3 and Figure 4 show the average results of the SRR metric per 4 different noise types (e.g., babble, airport announcements) and their standard deviation. These results directly support the objectives of the **Research Question 1**.

**Noise Hierarchy:** The results demonstrate a clear hierarchy in reconstruction effectiveness that correlates with the speech-like characteristics of each noise type. Speech-like conditions (babble and airport announcements) achieve significantly higher SRR values ( 9.5 dB) compared to non-speech noise types, with white noise showing the poorest performance ( 8 dB). This indicates that the auditory kernels are inherently better suited for representing structured, speech-like acoustic content.

**Denoising:** At higher SNR conditions (10 dB), substantial reconstruction quality is achieved using relatively few kernels (stops around 300 kernels/s for all cases), suggesting efficient capture of clean speech components. As SNR decreases,

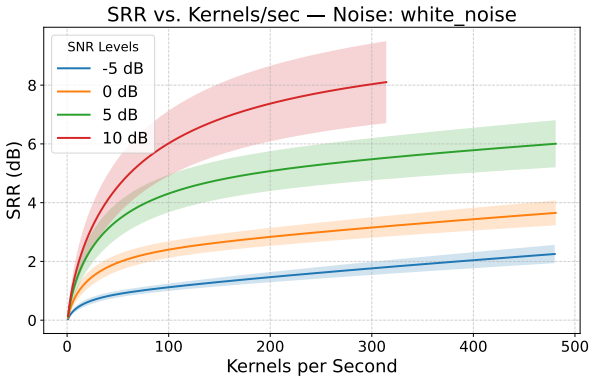



(a) Babble

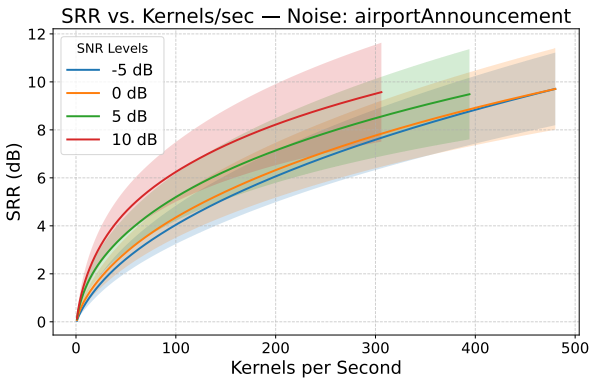



(b) Train Coming

Figure 3: Averaged SRR curves for (a) Babble and (b) Train Coming noise types.




(a) White Noise




(b) Airport Announcement

Figure 4: Averaged SRR curves for (a) White Noise and (b) Airport Announcement noise types.

more kernels are required to achieve similar quality, as the algorithm progressively captures noise components. This behavior demonstrates that auditory kernels naturally prioritize speech-like patterns in their kernel selection process. Which leads to the denoising effect of the auditory kernel reconstruction.

### 6.3 Expected Activations of Kernels Under Degradation

The kernel activation results presented here visualize the **normalized difference** in kernel usage between degraded speech and noise-only reconstructions. Each plot shows, for a given noise type at 5 dB SNR, how much more frequently a given kernel was activated during speech reconstruction compared to noise reconstruction. The Y-axis represents the *proportional difference* in usage (i.e., speech proportion minus noise proportion), and each bar corresponds to a kernel index.

- **Green bars** indicate kernels that were used more in speech reconstructions, suggesting they are *speech-preferred*.
- **Red bars** indicate kernels more active in noise, suggesting they are *noise-preferred*.

Figure 5 shows the difference pattern for *Train Coming* noise, Figure 6 for *Babble* noise, and Figure 7 for *White Noise*. The full set of plots, including raw noise-only kernel activations and other SNR levels, is included in Appendix B. These findings support the objective of **Research Question 3**.

Overall, the kernel activation behavior reveals a degree of specialization. Kernels that are commonly used to reconstruct speech tend to differ from those used for noise, particularly in unstructured noise conditions. Notably:

- **Structured, Non-speech-like Noise (Train Noise):** Some kernels shift roles depending on context. For instance, kernels #0 and #21 become noise-preferred even though it is highly a speech kernel based on other results. This is likely because it effectively models the rhythmic structure of mechanical noise, and it needs more kernels to replicate that waveform (y-label) since it is not similar to speech. That is why activations in the noise dominate activations in the speech.
- **Speech-like Noise (Babble, Airport):** In these cases, separation blurs. Speech-optimized kernels such as #0, #6, and #21 are activated for both signal and background, since both contain intelligible speech. This leads to certain kernels being labeled noise-preferred even though they also capture speech patterns, highlighting the ambiguity when background noise is linguistically structured.

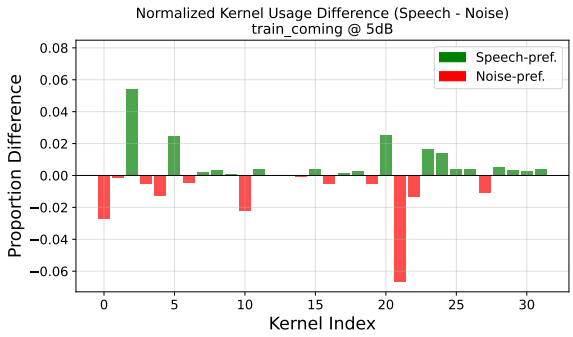



Figure 5: Normalized activation difference for Train Coming noise at 5 dB.

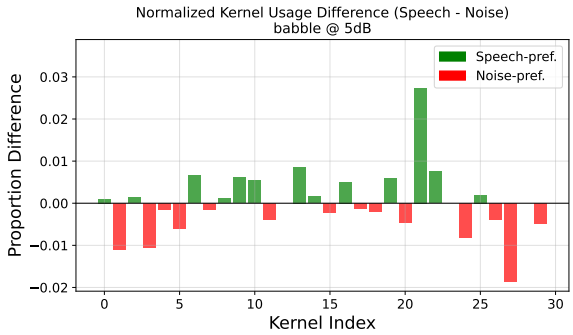



Figure 6: Normalized activation difference for Babble noise at 5 dB.

- **Unstructured Noise (White Noise):** The separation is most distinct. Kernel #15 consistently emerges as noise-preferred, while kernel #21 is speech-preferred. This indicates effective kernel specialization. Check Appendix B to see all the results.

# 7 Responsible Research

## 7.1 Ethical Considerations

This research utilizes speech data to analyze and reconstruct auditory signals using a biologically inspired sparse coding approach. Since speech data can sometimes carry sensitive personal information, we ensured that all audio samples used in this study came from a well-documented, publicly available dataset—MS-SNSD [12]. The dataset includes consented recordings from multiple speakers across diverse accents and is commonly used in speech processing research. No private or personally identifying information is present in the data, and all files are anonymized by speaker ID.

Furthermore, the use of synthetic degradation (e.g., adding background noise) rather than real-world surveillance or harvested data avoids the risk of reinforcing harmful or biased patterns. The focus of the research is on the perceptual and structural characteristics of signal reconstruction, not speaker identification or profiling, which helps mitigate ethical risks related to the misuse of voice data.

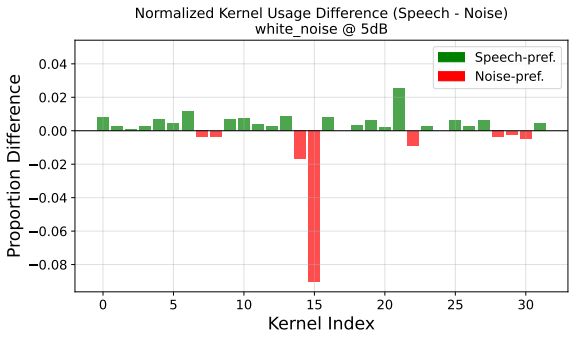



Figure 7: Normalized activation difference for White noise at 5 dB.

## 7.2 Reproducibility and Research Integrity

All major components of our research pipeline—including degrading, kernel-based Matching Pursuit encoding, reconstruction, and evaluation—have been implemented in Python using publicly available libraries [16; 20; 25]. The preprocessing, encoding, and evaluation procedures are modularized and documented to ensure that each step of the process can be replicated.

To support environmental reproducibility, particularly in perceptual quality assessments, the ViSQOL metric is run within a Docker container. This avoids dependency conflicts across systems and allows researchers to replicate results using the same controlled environment. Our experiments were executed on the DelftBlue cluster [26], and a batch-processing script was used to reconstruct all degraded samples efficiently. Parameters such as stopping condition (residual amplitude threshold of 0.1) are explicitly defined and fixed for consistency across experiments.

Although sparse coding methods like Matching Pursuit can be computationally intensive, especially on long sequences, our method demonstrates scalability through batch processing and selective encoding. We avoid overfitting to individual utterances by using a fixed, pre-trained kernel dictionary across all reconstructions.

In addition, large language models (LLMs) were used during the writing and implementation process to assist with grammar correction, wording clarity, and polishing the wording. LLMs were also used to generate or improve parts of the plotting and scripting code, such as figure formatting and automation routines. These uses were strictly supportive and did not influence the scientific interpretation or originality of the research design. The example prompts can be found in Appendix C

The code-base and experiment logs are publicly available to further support transparency and reproducibility; you can access the repository https://github.com/baturalpkars/RP_Auditory_Kernels [25]. Any findings from this research should be interpreted in the context of its design—focused on controlled degradation and single-speaker reconstruction—while keeping future real-world applications in perspective.

## 8 Discussion

This study explored whether auditory kernel-based sparse reconstruction can selectively preserve speech-like structure under degraded acoustic conditions and what types of patterns are prioritized by kernels. The results consistently support this hypothesis, revealing three key insights:

**1. Robust intelligibility and quality.** STOI and PESQ scores show that auditory kernels preserve intelligibility and perceptual quality well across all noise types and SNR levels. At high SNRs, the reconstruction even introduces a mild denoising effect, particularly for non-speech noise types such as white noise and train noise. Similar behavior was reported by Sigg et al. [11], who showed that sparse dictionaries can enhance noisy speech by suppressing noise atoms and emphasizing speech-relevant structures.

**2. Selective reconstruction behavior.** The SRR results demonstrate that speech-like noise (e.g., babble, airport announcements) is more easily captured by the auditory kernels, leading to higher SRR values with fewer activations. This suggests that the kernels are well-tuned to speech-like patterns. In contrast, for unstructured noise (e.g., white noise) or structured non-speech-like noise (e.g., train noise), SRR increases more slowly. However they achieve a good SRR on high SNRs with less kernels. This behavior supports the hypothesis that the reconstruction process implicitly prioritizes speech-like content. This aligns with observations by Mesgarani et al. [4], who found that cortical representations in the brain remain selective for structured inputs like speech, even in noisy environments.

**3. Kernel specialization emerges.** The kernel activation comparisons confirm that some kernels consistently activate for speech, while others are recruited more heavily during noise-only reconstructions. While specialization is clearest in unstructured noise settings, speech-like noise causes overlaps, leading to ambiguous kernel roles. This links auditory kernels with neurological findings by Souffi et al. [5], who found that some neurons consistently responded to speech even in noise, while others became more active for noise segments, indicating functional selectivity in the auditory system.

Together, these findings validate the biological inspiration behind auditory kernel models and provide evidence for their application in robust, explainable speech processing systems.

## 9 Conclusion and Future Work

This work investigated the effectiveness of auditory kernel-based sparse reconstruction under degraded speech conditions by addressing three core questions: (1) How much does the auditory kernel reconstruction selectively reconstruct speech-like patterns even in conditions where speech is degraded? (2) How does the quality of reconstructed signals evolve across different noise types and signal-to-noise ratio (SNR) in terms of perceptible quality and intelligibility? and (3) How can we quantify the selectivity of auditory kernel activations concerning different noise types and degraded speeches with those noises?

First, the SRR (Signal-to-Residual Ratio) curves showed that auditory kernels effectively capture signal energy even under noisy conditions. Reconstruction quality increased steadily as more kernels were used, especially for speech-like noise. At higher SNRs, a smaller number of kernels was sufficient to reconstruct speech, indicating efficiency. These findings suggest that the reconstruction process preserves meaningful speech structure even when the input is degraded, aligning with the strong intelligibility and perceptual scores observed across conditions (1).

Second, our results show that auditory kernels trained solely on clean speech remain effective even under significant noise. Reconstruction preserved intelligibility (STOI) and perceptual quality (PESQ) across various conditions, confirming that these kernels are robust to input degradation and remain biologically plausible (2).

Third, our kernel activation comparisons revealed consistent selectivity: some kernels strongly preferred speech, others were predominantly activated by noise. This selectivity was most pronounced in unstructured, low-SNR conditions and weaker in structured, speech-like noise, suggesting that sparse auditory representations naturally encode interpretable patterns (3).

While the results are promising, there are a few limitations to note. First, due to computational constraints—especially during large-scale batch processing on the Delft-Blue cluster—the ViSQOL metric was not included in the main results. Second, the dataset consists entirely of short, clean English speeches (3–6 seconds), which may not generalize to longer, conversational, or multilingual speech. Additionally, all reconstructions used a fixed pre-trained dictionary, which may limit adaptability to new acoustic domains or mixed speech scenarios.

For future work, a promising direction is to analyze individual kernel roles in more detail, identifying noise-sensitive units that could be dynamically suppressed during reconstruction, particularly in stationary or structured noise. Extending the kernel dictionary to cover overlapping speech and diverse acoustic scenes may also improve generalization. Moreover, because auditory kernels are biologically inspired, sparse, and computationally simple compared to phones or phonemes, they may contribute to improved automatic speech recognition (ASR) pipelines by acting as a biologically grounded, denoising layer that emphasizes and preserves the speech structure.

## References


[1] Michael S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, April 2002.

[2] Evan C. Smith and Michael S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, February 2006.

[3] Vivienne L. Ming and Lori L. Holt. Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, 126(3):1312–1320, September 2009.

[4] Nima Mesgarani, Stephen V. David, Jonathan B. Fritz, and Shihab A. Shamma. Mechanisms of noise robust representation of speech in primary auditory cor-

tex. *Proceedings of the National Academy of Sciences*, 111(18):6792–6797, May 2014.

[5] S. Souffi, C. Lorenzi, C. Huetz, and J.-M. Edeline. Robustness to Noise in the Auditory System: A Distributed and Predictable Property. *eneuro*, 8(2):ENEURO.0043–21.2021, March 2021.

[6] Nicole L. Carlson, Vivienne L. Ming, and Michael Robert DeWeese. Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus. *PLoS Computational Biology*, 8(7):e1002594, July 2012.

[7] Taishih Chi, Powen Ru, and Shihab A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, August 2005.

[8] Michael S. Lewicki. A signal take on speech. *Nature*, 466(7308):821–822, August 2010.

[9] Ismo Miettinen, Hannu Tiitinen, Paavo Alku, and Patrick Jc May. Sensitivity of the human auditory cortex to acoustic degradation of speech and non-speech sounds. *BMC Neuroscience*, 11(1):24, December 2010.

[10] Yuan Yao, Benshun Yi, and Yongqiang Yao. Speech Enhancement Under Aviation Noise. In *2006 International Conference on Wireless Communications, Networking and Mobile Computing*, pages 1–4, Wuhan, China, September 2006. IEEE.

[11] Christian D. Sigg, Tomas Dikk, and Joachim M. Buhmann. Speech enhancement with sparse coding in learned dictionaries. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4758–4761, Dallas, TX, USA, 2010. IEEE.

[12] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech 2019*, pages 1816–1820, 2019.

[13] Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization.

[14] Pavlos Papadopoulos, Andreas Tsiartas, James Gibson, and Shrikanth Narayanan. A supervised signal-to-noise ratio estimation of speech signals. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8237–8241, Florence, Italy, May 2014. IEEE.

[15] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, December 1993.

[16] D1mme. rp_auditory_kernels: Github repository. https://github.com/D1mme/rp_auditory_kernels, 2025.

[17] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[18] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752, Salt Lake City, UT, USA, 2001. IEEE.

[19] A.E. Conway. Output-based method of applying PESQ to measure the perceptual quality of framed speech signals. In *2004 IEEE Wireless Communications and Networking Conference (IEEE Cat. No.04TH8733)*, pages 2521–2526, Atlanta, GA, USA, 2004. IEEE.

[20] Miao Wang, Christoph Boeddeker, Rafael G. Dantas, and ananda seelan. ludlows/python-pesq: supporting for multiprocessing features, May 2022.

[21] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. ViSQOL: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, December 2015.

[22] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, September 2011.

[23] John Beerends, Andries Hekstra, Antony Rix, and M. Hollier. Perceptual evaluation of speech quality (pesq) - the new itu standard for end-to-end speech quality assessment - part ii - psychoacoustic model. *Journal of the Audio Engineering Society. Audio Engineering Society*, 50, 10 2002.

[24] Roy Patterson, Ian Nimmo-Smith, John Holdsworth, and Peter Rice. AN EFFICIENT AUDITORY FILTERBANK BASED ON THE GAMMATONE FUNCTION.

[25] Baturalp Karslıoğlu. auditory-kernel-reconstruction: Github repository. https://github.com/baturalpkars/RP_Auditory_Kernels, 2025.

[26] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 2). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2, 2024.

# A Full Perceptual Evaluation Plots

This appendix contains the complete set of PESQ and STOI evaluation plots across all noise types and SNR levels. These plots expand on the perceptual quality and intelligibility analysis presented in Section 6.1.

- PESQ plots: `pesq_avg_snr`$\{-5, 0, 5, 10\}$`.pdf`
- STOI plots: `stoi_avg_snr`$\{-5, 0, 5, 10\}$`.pdf`

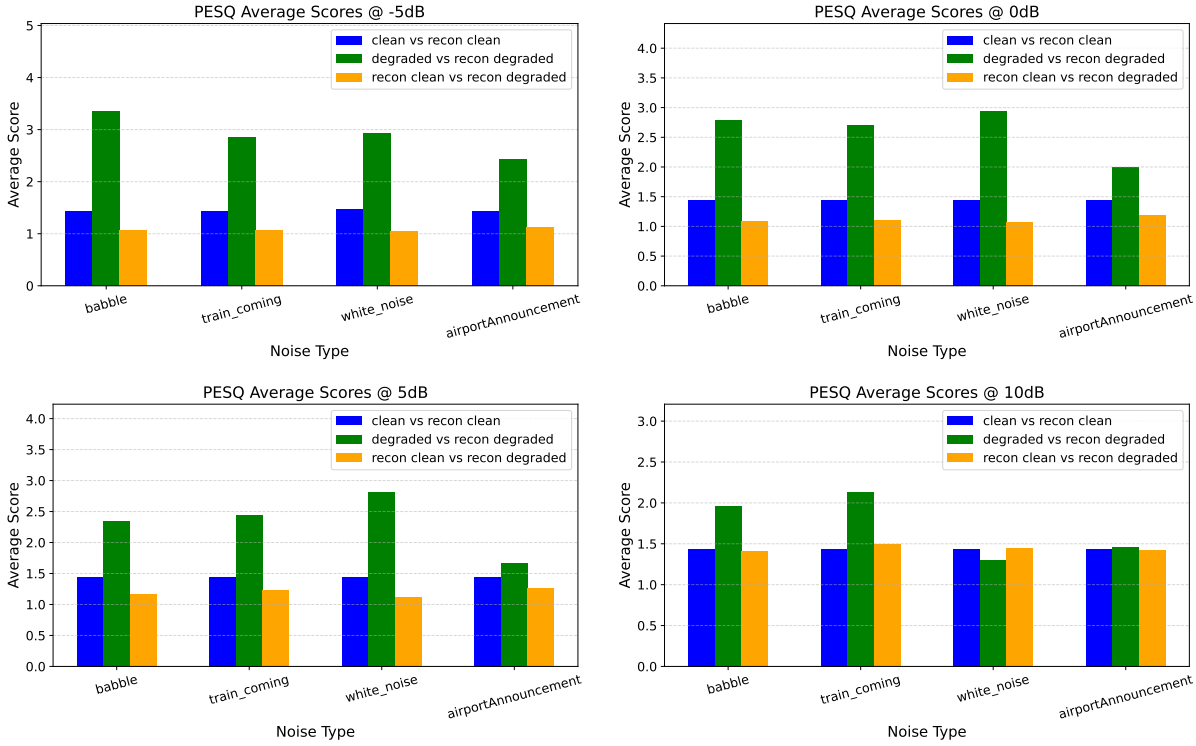



Figure 8: PESQ scores for each noise type across SNR levels.

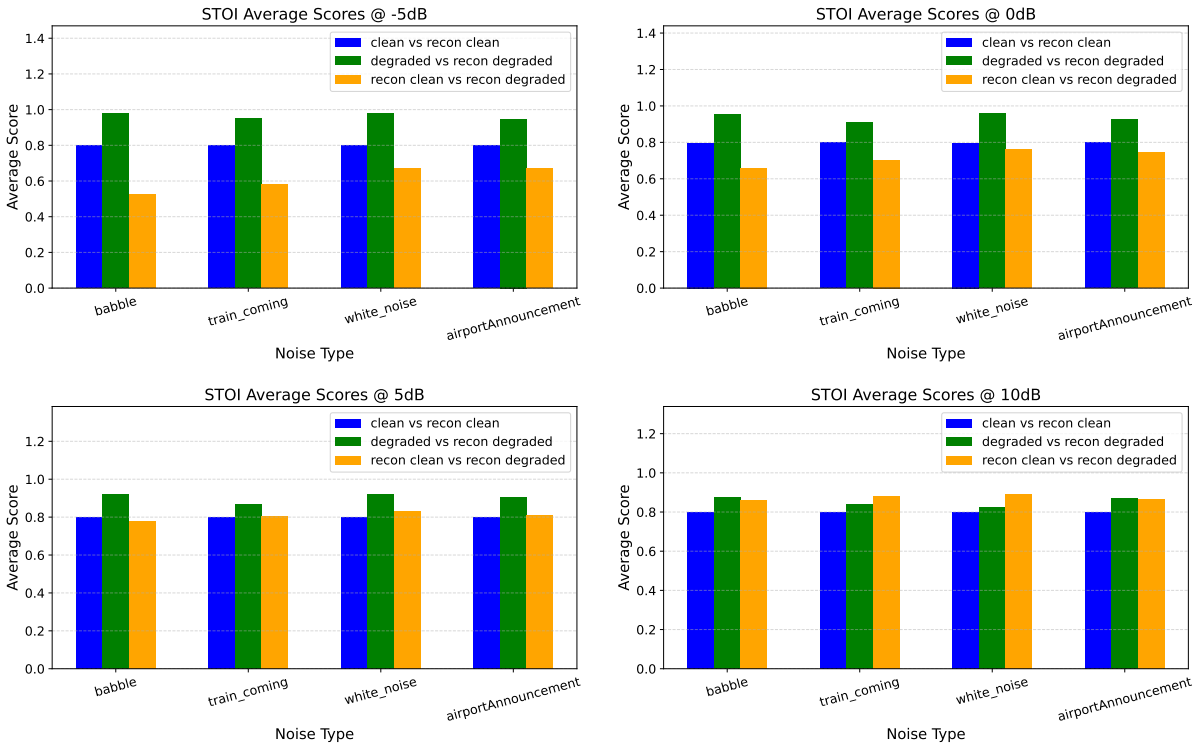



Figure 9: STOI scores for each noise type across SNR levels.

## B Kernel Activation Comparison Plots

This appendix contains the full set of kernel activation comparisons discussed in Section 6.3. For each noise type:

- Figure 10, 12, 14, and 16 show the raw kernel activation histograms during noise-only reconstruction for noise types; *babble, train coming, white noise, airport announcement*, respectively.
- (a)–(d) show the normalized proportional difference between degraded speech and noise kernel usage for SNR levels of $-5, 0, 5, 10$ dB.

Positive values in the difference plots indicate speech-preferred kernels; negative values indicate noise-preferred kernels.

### Babble Noise

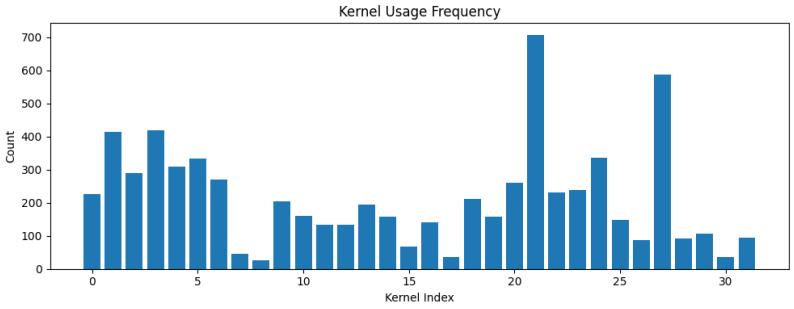
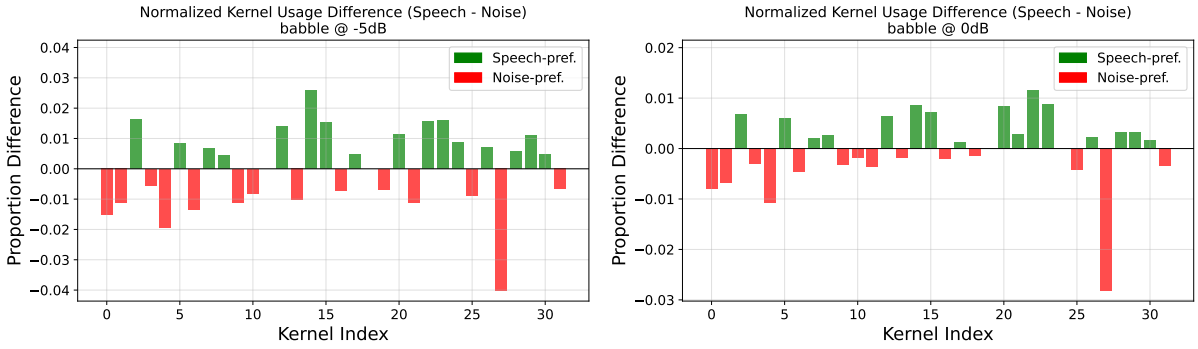


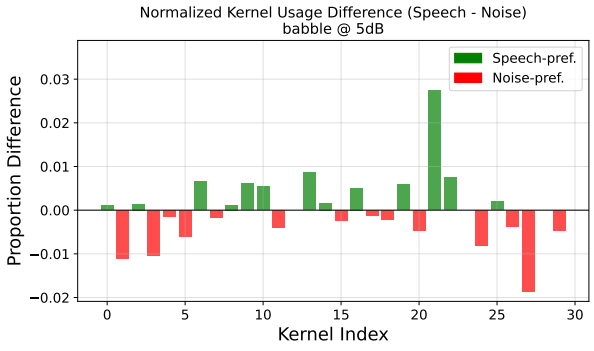

Figure 10: Noise-only kernel activations for **Babble**.



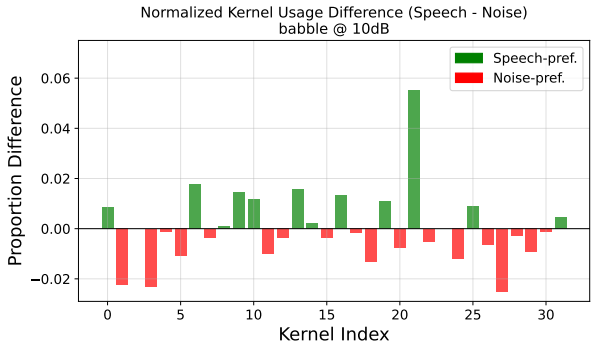

(a) SNR $= -5$ dB




(b) SNR $= 0$ dB




(c) SNR $= 5$ dB




(d) SNR $= 10$ dB

Figure 11: Kernel activation differences for **Babble** noise at different SNR levels.

## Train Coming Noise

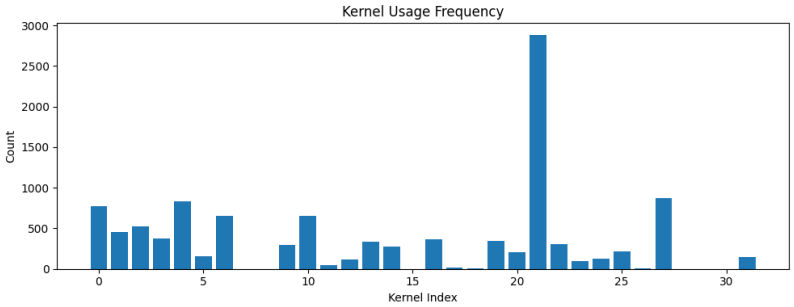
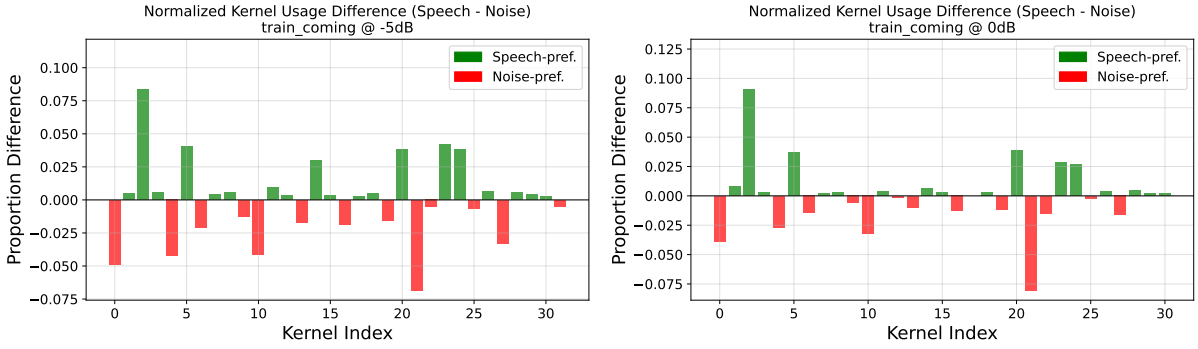


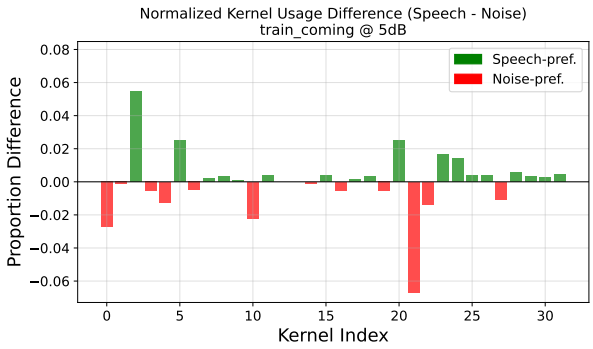

Figure 12: Noise-only kernel activations for **Train Coming**.



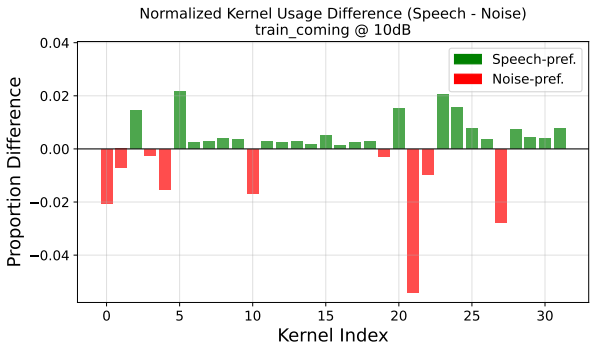

(a) SNR $= -5$ dB




(b) SNR $= 0$ dB




(c) SNR $= 5$ dB




(d) SNR $= 10$ dB

Figure 13: Kernel activation differences for **Train Coming** noise at different SNR levels.

## White Noise

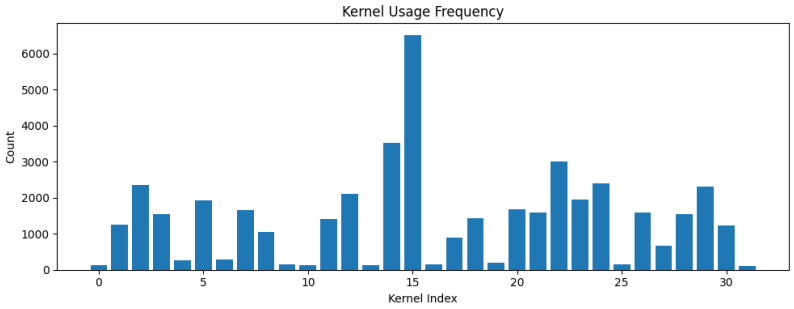
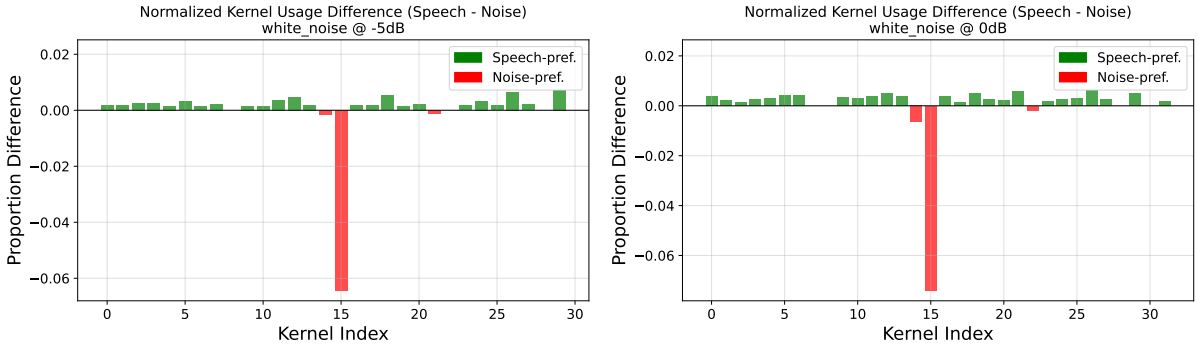


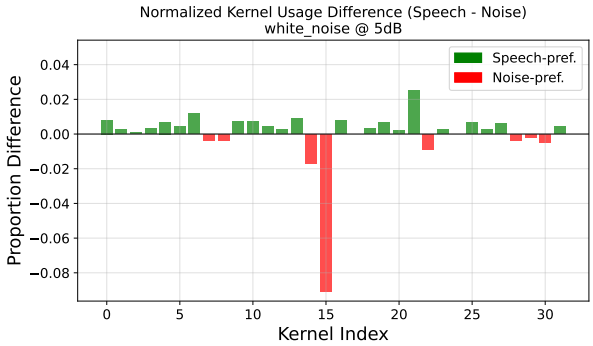

Figure 14: Noise-only kernel activations for **White Noise**.



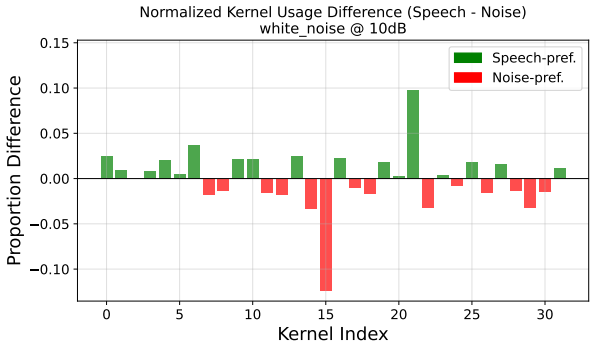

(a) SNR $= -5$ dB




(b) SNR $= 0$ dB




(c) SNR $= 5$ dB




(d) SNR $= 10$ dB

Figure 15: Kernel activation differences for **White Noise** at different SNR levels.

## Airport Announcement

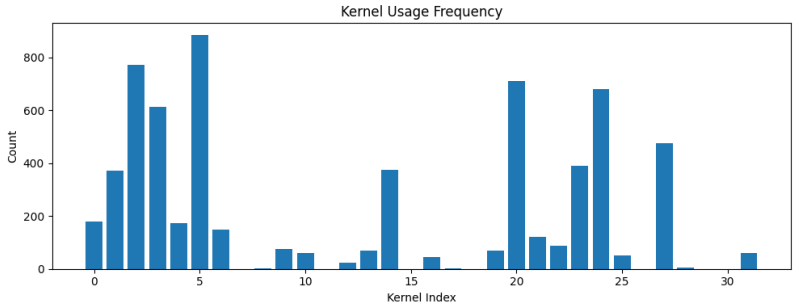
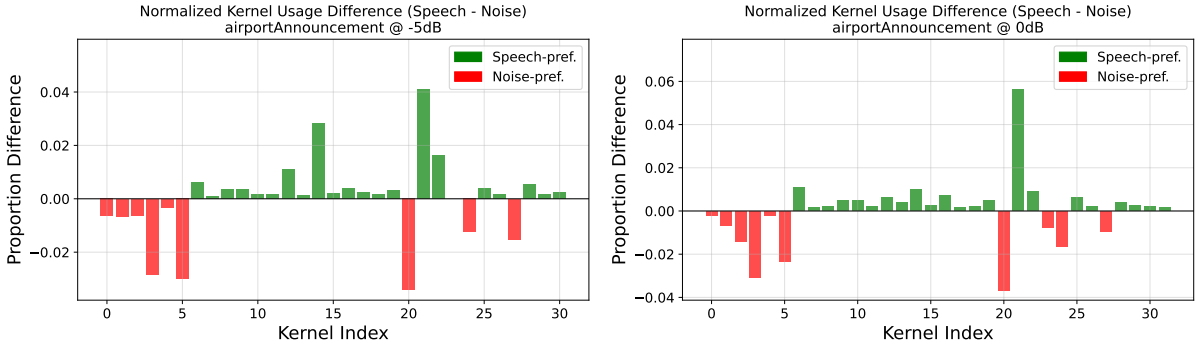


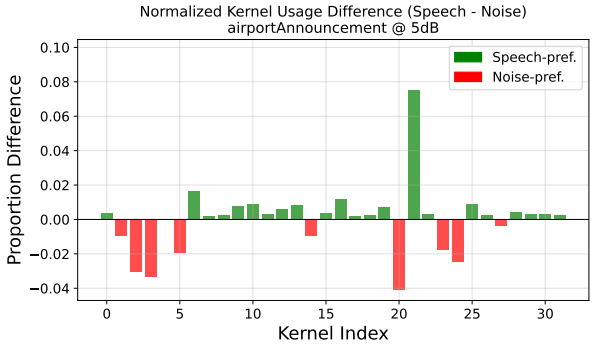

Figure 16: Noise-only kernel activations for **Airport Announcement**.



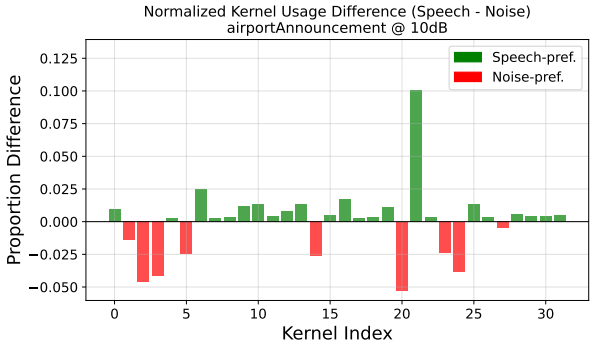

(a) SNR $= -5$ dB




(b) SNR $= 0$ dB




(c) SNR $= 5$ dB




(d) SNR $= 10$ dB

Figure 17: Kernel activation differences for **Airport Announcement** at different SNR levels.

## C Example Prompts and Responses for LLM Usage

To enhance the clarity and reproducibility of this research, we briefly document examples of Large Language Model (LLM) usage during writing and development. These examples illustrate how LLMs were used responsibly—for grammar polishing and code refactoring, not for generating novel scientific insights.

### Example 1: Prompt for Clarity and Academic Wording

**Prompt:**

> I'm writing a sentence for the discussion section of my research report. I want to describe that some kernels become more active when the noise is speech-like, which creates ambiguity during reconstruction. Can you help me rephrase the sentence to make it clearer and more academically sounding?
> Original sentence: "When the noise has speech characteristics, it activates similar kernels with speech and makes it harder to separate."

**LLM Response:**

> "Speech-like noise often activates similar kernels as those used for speech, leading to overlapping representations and increased ambiguity during reconstruction."

### Example 2: Prompt for Plotting Script Refactoring

**Prompt:**

> I'm creating bar plots in Matplotlib to compare average PESQ scores for four noise types at different SNRs. I want to combine two SNR conditions (0 dB and 10 dB) into one figure using subplots (top and bottom), and I want the legend to appear only once. Can you help me write or refactor the code accordingly?

**LLM Response (Python Snippet):**

```
fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(9, 8), sharex=True)
snr_levels = ['10', '0']

for idx, snr in enumerate(snr_levels):
    ax = axes[idx]
    for i, comp in enumerate(comparison_columns):
        ax.bar(x + i * bar_width, avg_data[comp], width=bar_width,
               label=comp.replace("_", " "), color=comparison_colors[comp])
    ax.set_title(f"PESQ Scores @ {snr} dB")
    ax.set_ylabel("Avg Score")

handles, labels = axes[0].get_legend_handles_labels()
fig.legend(handles, labels, loc='upper center', ncol=3, fontsize=11, frameon=True)

fig.tight_layout(rect=[0, 0, 1, 0.95])
fig.savefig("combined_pesq_plot.pdf")
```