# Auditory Kernels for Representing Degraded Speech

Baturalp Karslioglu [1], Supervisor: Dimme De Groot [1], Professor: Jorge Martinez [1]

[1]TU Delft – CSE3000 Research Project

## Background & Motivation

Auditory kernels are biologically inspired filters that mimic how the human auditory system—particularly the cochlea and cortex—encodes sound. Lewicki showed that such filters emerge from sparse, efficient representations of natural sounds, and later work demonstrated their effectiveness in modeling human speech perception [1, 2, 3].

This study examines how these auditory kernels behave under realistic degradations. Therefore, Train Station has chosen as the simulated environment. We analyze the signals that are reconstructed by auditory kernels to assess (1) intelligibility and quality (perceptibility), (2) reconstruction process with Signal-to-residual ratio (SRR) and (3) which kernels are activated in response to speech vs. noise—providing insight into how structure is preserved under noise.

## Research Questions

- **RQ1:** 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)?*
- **RQ2:** 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?*
- **RQ3:** 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?*

## Methodology

**Auditory Kernels:**

We use 32 biologically-inspired auditory kernels trained via sparse coding on the TIMIT corpus [4]. These kernels capture speech-like spectrotemporal patterns, similar to receptive fields in the human auditory system [1, 2].

**Reconstruction Framework:**

Matching Pursuit iteratively selects kernel matches to reconstruct the signal, terminating when the maximum inner product between the residual and any kernel drops below 0.1. This ensures consistent reconstructions and prevents overfitting [5].

**Dataset:**

- **Clean speech signals** consists of 25 Male, 25 Female speakers with 2 speech each, end up in 100 speech samples from Microsoft Scalable Noisy Dataset (MSND) [6].
- **Noise types:** 4 real-world background noises— *babble, airport announcement [6], train arrival, and white noise* to replicate the Train Station environment— added to clean signals at SNR levels of -5, 0, 5, and 10 dB to simulate degraded conditions.

**Evaluation Metrics:**

We assess the reconstruction using:

- **PESQ** [7, 8]: Models perceptual quality by comparing the internal auditory representations of reference and degraded signals.
- **STOI** [9]: Estimates speech intelligibility based on short-time spectral correlations between clean and reconstructed signals.
- **SRR (dB)**: Signal-to-Residual Ratio tracks reconstruction fidelity across kernel additions, showing how efficiently structure is captured.
- **Kernel activation histograms**: Reveal kernel selectivity by comparing normalized activations of kernels across reconstructed speech and noise-only inputs.

## Responsible Research

All speech samples come from a public, consented dataset (MS-SNSD) with no identifying information [6]. We simulate degradation with synthetic noise and do not perform speaker recognition to avoid ethical risks.

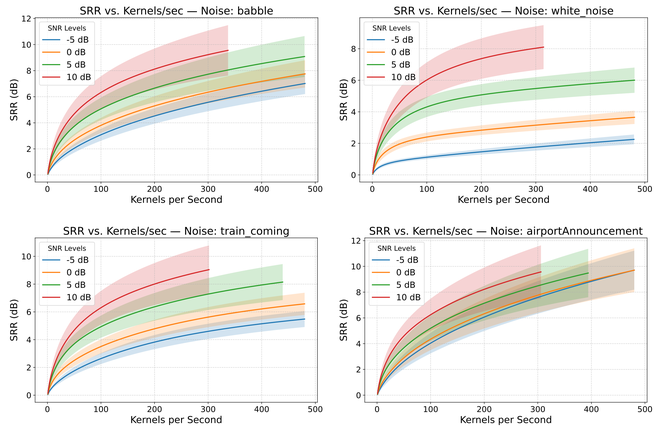
The pipeline is fully implemented in Python with public libraries [4], and accessible through GitHub [10]. All steps—from degradation to reconstruction and evaluation—are accessible and reproducible.

## RQ 2: Perceptual Quality (PESQ) & Intelligibility (STOI) score


Observations:

- Higher SNRS, reconstructed performs better than or same with the baseline. Shows denoising.
- Speech-like noise types (e.g., babble, airport) reduce intelligibility more than non-speech-like noise types (e.g., white, train).
- Auditory kernel-based reconstructions retain quality and intelligibility even in degraded conditions.

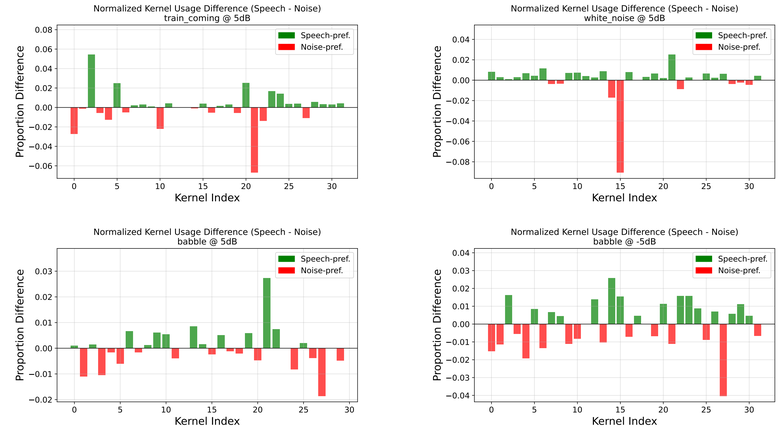
## RQ1: Reconstruction Efficiency (SRR vs. Kernel Rate)




Observations:

- Speech-like noise (e.g., babble) reconstructed better (SRR) than other noise-types, shows kernels capture speech patterns.
- At higher SNRs we use less kernels to achieve similar reconstruction, shows an evidence for denoising (SNR = 10db stops at 300 kernels/sec).

## RQ3: Kernel Selectivity (Speech vs. Noise)




Interpretation:

- Auditory kernels exhibit structured selectivity: some respond more to speech, others to noise.
- Speech-like noise (e.g., babble) overlaps with speech kernel usage, reducing separation clarity (#0, #6, #21).
- Distinct, non-speech noise (e.g., white) activates separate kernel sets, making it easier for the separation.

## Conclusion and Future Work

- **Robustness:** Auditory kernels trained on clean speech effectively reconstruct intelligible and perceptible (preserves quality) speech under various noise types and SNR levels. Aligns with Sigg's work [11].
- **Efficiency:** Fewer kernels are needed at higher SNRs (denoising); SRR curves shows that reconstruction preserves the speech patterns. Aligns with Mesgarani's work [12].
- **Selectivity:** Kernel activations distinguish speech from noise—especially for unstructured noise—revealing interpretable structure. Aligns with Souffi's work [13].
- **Limitations:** Experiments use short English samples and exclude ViSQOL (another perceptibility metric) due to computational limits.
- **Future Work:** Extend to multilingual speech, longer contexts, and dynamic kernel adaptation in ASR pipelines for speech enhancement.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[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] Nima Mesgarani, Stephen V. David, Jonathan B. Fritz, and Shihab A. Shamma. Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences*, 111(18):6792–6797, May 2014.

[13] 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.