# Scaling Auditory Kernel Dictionaries: The Sparsity-Fidelity Trade-off in Speech Reconstruction

Prarthana Badiger, Dimme de Groot, Jorge Martinez

Faculty of EEMCS, Delft University of Technology

## Introduction

**Efficient Coding Hypothesis:** Sensory systems minimise neural resources while maximising information transfer [1].

**Auditory kernels:** learned waveform templates whose shapes are optimised to capture the statistical structure of sound (analogous to the impulse responses of auditory nerve fibres)

**The Signal Representation (Reconstruction)**

To mathematically describe how an acoustic waveform is reconstructed from discrete spikes, Smith and Lewicki (2006) define the signal x(t) as a linear superposition of the kernel functions.

$$x(t) = \sum_m \sum_i s_i^m \phi_m(t - \tau_i^m) + \epsilon(t)$$

- $\phi_m$ represents the set of different kernel functions ( waveforms)
- $\tau_i^m$ is the precise temporal position of the i-th instance of a kernel
- $s_i^m$ is the analog coefficient (the amplitude or "spike")
- $\epsilon(t)$ is the additive noise, which represents the final residual error left over after reconstruction

**The Matching Pursuit Algorithm (Encoding)**

Non-linear, iterative algorithm used to decompose and encode a complex signal into a linear expansion of waveforms selected from a redundant dictionary of kernel functions. [2]

**Gradient Ascent (The Learning Algorithm)**

Optimization algorithm used during the training phase to dynamically update the shapes and lengths of the kernel functions so they can encode sounds more efficiently. [1]

## Research Question

**Research gap**

It is unknown how scaling beyond 32 kernels affects coding efficiency

***How does increasing the number of learned auditory kernels beyond 32 impact the trade-off between coding sparsity and reconstruction fidelity in human speech?***

Sub questions:

- **RQ1.** How does scaling the number of kernels to 64, 128, and 256 affect the rate-fidelity curves (kernel activations per second vs.\ Signal-to-Reconstruction Ratio in dB), and do any objective improvements translate to perceptible differences in speech quality?
- **RQ2.** Do the newly learned kernels in expanded dictionaries maintain the biologically accurate asymmetric structures observed in the 32-kernel dictionary?
- **RQ3.** What is the impact of larger dictionary sizes on kernel utilisation, as measured by spectral coverage and pairwise redundancy?

References

[1] Evan C. Smith and Michael S. Lewicki. Efficient auditory coding. Nature, 439(7079):978– 982, 2006.
[2] Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing, 41(12):3397–3415, 1993.

## Methodology

**Data Preparation:** Audio is normalized, bandpass filtered (100–6000 Hz), and sampled at 16 kHz using the TIMIT corpus.

**Dictionary Training:** Four dictionary sizes (32, 64, 128, 256 kernels) initialised from Gaussian noise; baseline (6 epochs) and optimised (stricter stopping condition, kernel dropout, decaying step size, 8 epochs)

**Encoding & Evaluation:** Matching Pursuit encodes speech into sparse spikes; rate-fidelity curves extracted (activations/s vs. SRR in dB)

**Biological Comparison:** Learned kernels compared to 755 cat auditory nerve fibre revcor filters via cross-correlation

**Dictionary Utilization Analysis:** Spectral coverage across 7 frequency bands; pairwise redundancy via cross-correlation between all kernel pairs

**Perceptual Evaluation:** MUSHRA listening test (30 participants) + PESQ objective speech quality scores

## Findings

### 128 kernels is the only size that improves reconstruction (RQ1)

- **Without optimisation**: all sizes converge to ~16.5 dB - no benefit from more kernels
- **With optimisation**: 128 kernels achieves 20.15 dB (+3.6 dB over baseline)
- **256 kernels**: no improvement regardless of training strategy
- **Critical factor**: stricter stopping condition forces kernels to learn finer residual structures

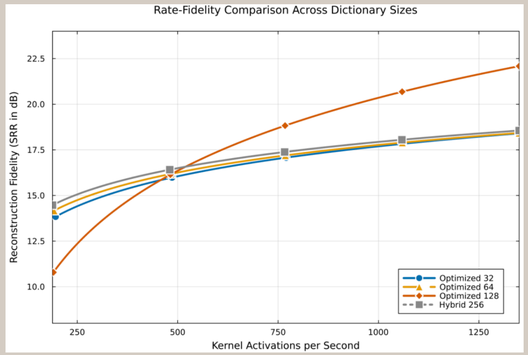


**Fig. 1**: Rate-fidelity curves for the best-performing model at each dictionary size. Each point is one audio file. The optimised 128-kernel model (orange) separates from all others above 1,000 activations/s.

### No perceptible quality difference at matched bit rates (RQ1)

- **Equal bit rate comparison** at 23.4 kbps across all four dictionary sizes
- All models scored 38–41 ("poor" quality range)
- **PESQ wideband scores**: 1.45–1.55, confirming no perceptual advantage

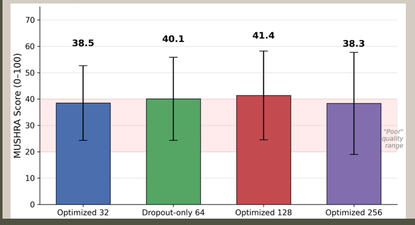


**Fig. 2**: Mean MUSHRA scores with standard deviation (n=30, 23.4 kbps). All models fall within the "poor" quality range (20–40) with no significant differences.

### Biological resemblance preserved across all scales (RQ2)

- **Top-20 correlations stable**: r = 0.81 (32), r = 0.80 (64), r = 0.82 (128)
- **Population-wide correlation decreases**: 0.72 → 0.61 → 0.39
- Decline reflects lack of sufficient training (only trained for two epochs)

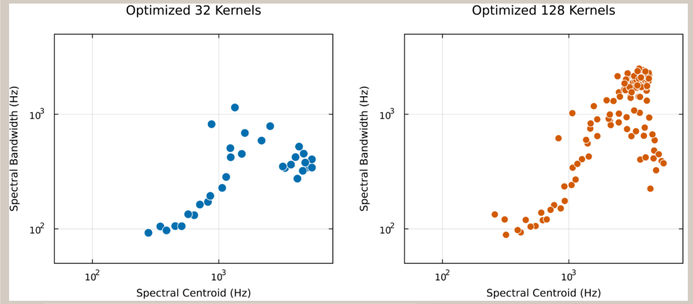


**Fig. 3**: Spectral centroid vs. bandwidth for the 32-kernel (left) and 128-kernel (right) dictionaries.

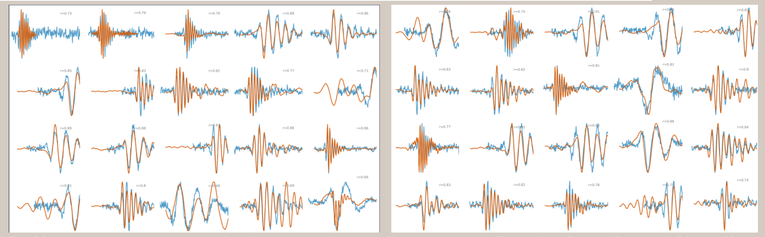

**Fig. 4:** Top 20 learned kernels (orange) overlaid with best-matching cat auditory nerve revcor filters (blue) for the 32-kernel (left) and 128-kernel (right) dictionaries.

### Spectral imbalance explains the 256-kernel failure (RQ3)

- **All kernels active at every dictionary size** - no dead kernels
- **128**: most balanced spectral coverage (10-28 kernels per band)
- **256**: 37% of kernels concentrated in 3-4 kHz band; low frequencies under-covered
- Gradient ascent concentrates kernels where speech energy is highest rather than spreading across the spectrum

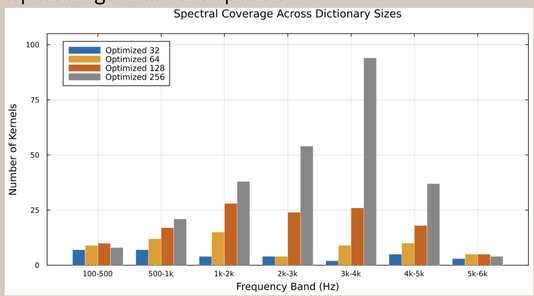


**Fig. 5**: Distribution of kernel spectral centroids across seven frequency bands. The 128-kernel dictionary achieves balanced coverage; the 256-kernel dictionary concentrates 37% of kernels in 3–4 kHz.

## Conclusions

- 128 kernels is the practical upper bound for greedy Matching Pursuit on speech
- Objective SRR gains do not translate to perceptual gains at matched bit rates
- Biological resemblance preserved across all dictionary scales
- 256 fails due to spectral imbalance, not inactive kernels

**Limitations:**

- Coding cost measured in activations/s, not entropy-based bits/s
- Hyperparameters not exhaustively searched for 256 kernels
- MUSHRA evaluates quality, not speech intelligibility