Efficient Auditory Coding Under Reverberation

Effects of Reverberation on Sparse Auditory Kernel Representations of Speech

Supervisor: Ir. Dimme de Groot [1] Author: Andrei-Rares Patrut [2] Responsible Professor: Dr. Jorge Martinez [2]

[1]EEMCS, Delft University of Technology

Background and Motivation

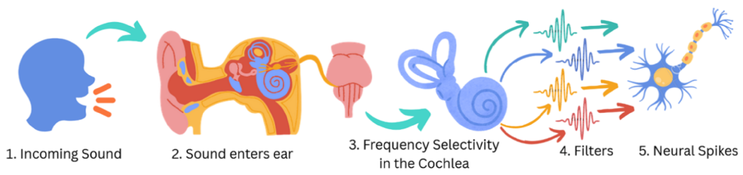


Figure 1. High-level illustration of auditory processing, from incoming sound to neural representations.

The efficient coding hypothesis suggests that sensory systems represent sound using compact and efficient internal codes. Sparse auditory-kernel models apply this idea to speech by reconstructing signals from a small number of learned waveform components [1].

However, real-world speech is rarely clean. In rooms, the direct sound is mixed with delayed reflections, producing reverberation and temporal smearing. This can be modeled as

$$y(t) = x(t) * h(t),$$

where $x(t)$ is clean speech and $h(t)$ is the room impulse response [2].

While sparse auditory coding has been studied extensively for clean speech, it is less clear how reverberation affects the learned kernels and the resulting sparse representation. This project investigates *how reverberation changes kernel structure, reconstruction sparsity, objective reconstruction quality and intelligibility, and perceived speech quality*.

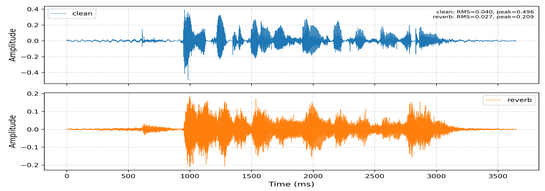


Figure 2. Example clean and reverberant speech waveform. Reverberation spreads speech energy over time.

Research Questions

RQ1. How do auditory kernels trained on clean speech differ from auditory kernels trained on naturally reverberant speech in terms of waveform shape, duration, spectral centroid, spectral spread, and dictionary similarity?

RQ2. How does reverberation affect the number of kernel activations required to reconstruct speech under fixed Matching Pursuit settings?

RQ3. How do clean-trained and reverberation-trained kernels differ in reconstruction performance for clean, naturally reverberant, and artificially reverberant speech, measured using SRR, STOI, PESQ, and subjective listening scores?

Method and Experimental Setup

Auditory Kernels:

- We compare 32 learned auditory kernels trained on clean speech and naturally reverberant speech.
- Speech is approximated as a sparse sum of kernels:

$$s(t) \approx \sum_{k=1}^{K} \alpha_k \phi_{f(k)}(t - \tau_k) + \epsilon(t).$$

Reconstruction Framework:

- Matching Pursuit is used for sparse reconstruction, with either amplitude-threshold or fixed-budget stopping depending on the experiment [3].
- The number of kernel activations measures sparsity, while the residual measures reconstruction fidelity.

Dataset:

- Clean speech was taken from TIMIT, an American English speech dataset sampled at 16 kHz [4].
- Natural reverberant speech was recorded by replaying clean TIMIT utterances in the TU Delft INSYGHT Lab using 2 loudspeakers and 4 microphones.
- The measured room reverberation time was $T_{30} = 0.76 \pm 0.03\,\mathrm{s}$, also used for artificial reverberation.
- D_{clean} was trained on clean speech, while D_{reverb} was trained on naturally reverberant speech.

Evaluation Metrics:

- **Activations/s** measures how many kernel activations are needed per second of speech. Lower values indicate a more compact sparse reconstruction.
- **SRR** measures waveform reconstruction fidelity from the residual error. Higher SRR means the reconstruction explains more of the input signal.
- **STOI** estimates speech intelligibility by comparing short-time spectral structure with the clean reference [5].
- **PESQ** estimates perceived speech quality by comparing reconstructed speech with the clean reference [6].
- **WebMUSHRA** provides subjective 0–100 quality ratings from 20 listeners [7].

RQ1: Learned Dictionaries

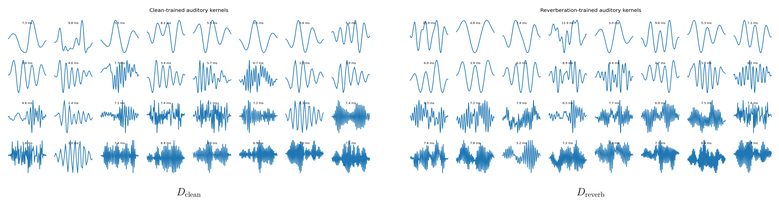


Figure 3. Clean-trained and reverberation-trained auditory kernels, sorted by spectral centroid.

RQ1: Kernel Changes Under Reverberation

Training on reverberant speech changed the learned dictionary structure. The reverberation-trained kernels became longer, shifted toward lower spectral centroids, and showed broader spectral spread. This suggests that the dictionary adapts to the temporal smearing and altered time-frequency statistics introduced by room reflections.

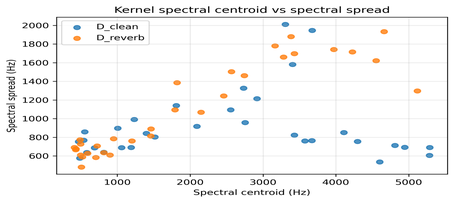


Figure 4. Spectral centroid and spread of kernels in both dictionaries.

Measure	D_{clean}	D_{reverb}	Change
Mean duration	5.64 ms	7.28 ms	+29%
Max duration	9.81 ms	15.94 ms	+62%
Mean centroid	2524 Hz	1934 Hz	−23%
Mean spread	913 Hz	1085 Hz	+19%
Peak frequency	2606 Hz	2323 Hz	−11%

Table 1. Main feature differences between clean-trained and reverberation-trained kernels.

Observations:

- Longer kernels are consistent with reverberant energy being spread over time.
- Lower centroids and peak frequencies suggest a shift toward lower-frequency structure.
- Larger spectral spread indicates that D_{reverb} covers broader frequency regions.
- The mean one-to-one similarity was 0.719, with 21/32 matched kernels above 0.7.
- Reverberation modifies the dictionary, but many clean-speech kernel structures remain preserved.

RQ2: Reverberation Reduces Sparse Reconstruction Efficiency

Reverberation increased the number of Matching Pursuit activations needed to reach the same amplitude-based stopping criterion. This means that reverberant speech was less sparse than clean speech under the same reconstruction settings.

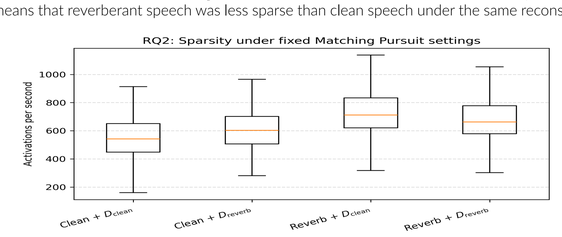
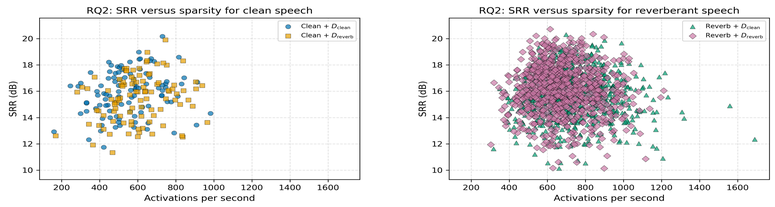


Figure 5. Activation rates for clean and reverberant inputs reconstructed with D_{clean} and D_{reverb}.



(a) Clean input (b) Naturally reverberant input

Figure 6. SRR versus activation rate for RQ2. Reverberant speech requires more activations than clean speech, while D_{reverb} partly reduces the activation rate for reverberant input.

Condition	Activations/s	SRR (dB)
Clean + D_{clean}	553.2 ± 155.8	15.85 ± 1.69
Clean + D_{reverb}	604.8 ± 154.1	15.49 ± 1.74
Reverb + D_{clean}	733.0 ± 166.8	15.81 ± 1.76
Reverb + D_{reverb}	681.6 ± 149.5	16.12 ± 1.85

Table 2. RQ2 sparsity and reconstruction fidelity under the same Matching Pursuit stopping criterion.

Observations:

- Reverberant speech required more activations than clean speech, showing reduced sparsity.
- With D_{clean}, the activation rate increased from 553.2 to 733.0 activations/s.
- For reverberant input, D_{reverb} reduced the rate from 733.0 to 681.6 activations/s.
- The matched dictionary gave slightly higher SRR in each acoustic condition.
- D_{reverb} partly compensates for reverberation, but does not make reverberant speech as sparse as clean speech.

RQ3: Reconstruction Performance

RQ3 evaluates whether the reverberation-trained dictionary improves reconstruction performance. The results show an important distinction: D_{reverb} can better reconstruct the reverberant waveform, but this does not necessarily improve intelligibility or perceived quality.

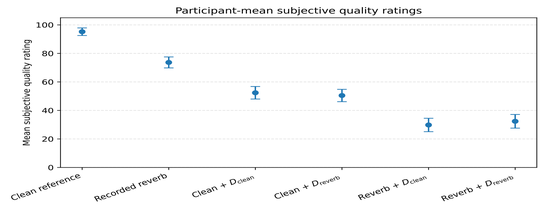


Figure 7. Participant-mean subjective quality ratings. Error bars indicate 95% confidence intervals.

Condition	SRR (dB)	STOI	PESQ
Clean + D_{clean}	17.26	0.948	1.71
Clean + D_{reverb}	16.36	0.939	1.55
Natural + D_{clean}	15.17	0.654	1.26
Natural + D_{reverb}	16.17	0.654	1.26
Artificial + D_{clean}	14.10	0.628	1.21
Artificial + D_{reverb}	14.01	0.621	1.20

Table 3. Objective reconstruction metrics.

Observations:

- For clean speech, D_{clean} performed best on SRR, STOI, PESQ, and subjective quality.
- For natural reverberation, D_{reverb} improved SRR from 15.17 to 16.17 dB.
- STOI and PESQ remained almost unchanged for natural reverberation, despite the SRR gain.
- Subjective ratings followed the same pattern: D_{reverb} was slightly preferred for reverberant input, but the difference was modest.
- Artificial reverberation did not benefit from D_{reverb}, suggesting that the learned adaptation is specific to the natural recording condition.

Main Takeaway

Reverberation changes sparse auditory representations, but the benefit of reverberation-trained kernels depends on the evaluation criterion.

- **RQ1:** Reverberation-trained kernels are longer, lower in spectral centroid, and broader in spectral spread.
- **RQ2:** Reverberant speech is less sparse and requires more kernel activations for reconstruction.
- **RQ3:** D_{reverb} improves SRR for natural reverberation, but this does not clearly improve STOI, PESQ, or subjective quality.
- Overall, D_{reverb} improves sparse reconstruction fidelity more than perceptual speech quality.

Limitations and Future Work

Limitations:

- The natural reverberant data came from one recording room and one loudspeaker–microphone setup.
- The subjective test used 20 participants and 16 trials, so perceptual conclusions should be interpreted cautiously.
- The results depend on the chosen Matching Pursuit stopping criteria and dictionary size.

Future work:

- Test multiple rooms, reverberation times, and source–receiver positions.
- Train dictionaries on more diverse reverberant conditions to improve generalization.
- Compare Matching Pursuit with other sparse approximation methods.
- Evaluate whether reverberation-aware kernels can support speech enhancement or automatic speech recognition.

References

[1] Evan C. Smith and Michael S. Lewicki. "Efficient Auditory Coding". In: *Nature* (2006).
[2] Jont B. Allen and David A. Berkley. "Image Method for Efficiently Simulating Small-Room Acoustics". In: *The Journal of the Acoustical Society of America* (1979).
[3] Stéphane G. Mallat and Zhifeng Zhang. "Matching Pursuits with Time-Frequency Dictionaries". In: *IEEE Transactions on Signal Processing* (1993).
[4] John S. Garofolo et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus". In: (1993).
[5] Cees H. Taal et al. "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* (2011).
[6] Antony W. Rix et al. "Perceptual Evaluation of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Networks and Codecs". In: 2001.
[7] Michael Schoeffler et al. "webMUSHRA – A Comprehensive Framework for Web-Based Listening Tests". In: *Journal of Open Research Software* (2018).