Perceptual Modeling Through an Auditory-Inspired Sparse Representation

EPFL-Idiap-ETH Sparsity Workshop 2015

Raphael Ullmann¹,² and Hervé Bourlard¹,²
¹Idiap Research Institute and ²EPFL
Prediction of Perceived Noise Intrusiveness

- Clean Speech Signal
- Noise
- Test Speech Signal
- Telecommunication System
- Mean Opinion Score (MOS)
Prediction of Perceived Noise Intrusiveness

- 10 real-world noise types with signal-to-noise ratios (SNR) at 3–40 decibels
- 3 datasets (500+ recordings)
- Noise intrusiveness ratings from listeners, scored on a 5-point scale
Prediction of Perceived Noise Intrusiveness

Clean Speech Signal + Noise → Telecom System → Test Speech Signal

Objective Quality Measure → Objective (Predicted) Mean Opinion Score

Mean Opinion Score (MOS)
Why Sparsity?

• Traditional approach
  • Combine acoustic features (noise level, variance, spectral composition)

• Our study: Focus on low-level sensory coding principles
  • Efficient Coding Hypothesis:
    “(...) our perceptions are caused by the activity of a rather small number of neurons selected from a very large population (...)” — [Barlow, 1972]
  • Redundancy reduction to help make sense of sensory inputs [Olshausen & Field, 1996]
Efficient Auditory Coding — Model

- Generative waveform model [Lewicki & Sejnowski, 1999]:
  \[ \hat{x}(t) = \sum_{m=1}^{M} \sum_{i=1}^{I_m} \alpha^i_m \phi_m(t - \tau^i_m) \]

- Shiftable kernels \( \{ \phi_1(t), \ldots, \phi_m(t), \ldots, \phi_M(t) \} \), can have different lengths

- Use Matching Pursuit to approximate \( \hat{x} = \Phi \alpha \), includes translation of kernels

- May think of each kernel instance as a population of spiking auditory neurons → “Spike Coding”

---

• How to choose the dictionary $\Phi = \{\phi_1(t), \ldots, \phi_m(t), \ldots, \phi_M(t)\}$?

Learn a dictionary from natural environmental noises [Smith & Lewicki, 2006]

Figure 1 illustrates the spike code model and its efficiency in representing a natural sound with the use of spikes. Each spike (oval) represents the temporal position and centre frequency of a segment of the word ‘canteen’ (input). The coloured arrows and curves indicate the correspondence between the spikes and the underlying acoustic structure represented by the kernel functions. Alignment of the spikes with respect to the kernels is accurate with little residual error (reconstruction and residual).

Figure 2a shows the learned kernel functions (red curves) for the natural sounds ensemble. To optimize the kernel functions we derived a code to an ensemble of sounds to which the auditory system is adapted to an unknown mixture of three broad categories of natural sounds. The kernel functions were optimized with cat physiological data (small blue dots) and with kernel functions of vowel alone (green triangles). They also adapt in temporal extent, with longer and shorter functions.

Figure © 2006 Nature Publishing Group

Perceptual Model — Dictionary

- Use a dictionary of analytically defined auditory filter shapes ("gammatones")

- We use 32 gammatones sampled at 16 kHz, generated with Slaney’s toolbox

---

Perceptual Model — Noise Signal Analysis

Kernel instances $\phi_{j(k)}(t)$ are localized in time and frequency.

Kernel instances are called atoms or “spikes”.
Perceptual Model — Noise Signal Analysis

Kernel instances $\phi_{j(k)}(t)$ are localized in time and frequency.

Kernel instances are called atoms or “spikes”.

Compute number of spikes (i.e., $\ell_0$ norm)
Perceptual Model — Noise Signal Analysis

Kernel instances $\phi_{j(k)}(t)$ are localized in time and frequency.

Kernel instances are called atoms or “spikes”.

Get $\ell_0$ norm over time.
Take the 5$^{th}$ percentile.
Perceptual Model — Evaluation

5th percentile of “spikes” over time highly correlates with subjective scores of noise intrusiveness.
Why Does It Work? — Because of Greedy Pursuit

- Decrease of spike energies (black line) depends on signal type
  - White noise is a kind of “worst case”, i.e., it does not correlate well with any kernel in the dictionary
- Logarithmic changes in sound energy produce linear changes in spike counts
- Greedy decomposition captures high-energy sounds first

![Graph showing energy of spikes/residual vs MP iteration for White Noise]
Why Does It Work? — Because of the Dictionary

Some tests with narrowband noises

![Graph showing Equivalent Spike Increment vs. Noise CF and Noise BW]
Why Does It Work? — Because of the Dictionary

Some tests with narrowband noises

- ERB-wide noise at varying center frequencies (CF)

→ Spike count similar to noise weighting curves
Why Does It Work? — Because of the Dictionary

Some tests with narrowband noises

- ERB-wide noise at varying center frequencies (CF)
  - Spike count similar to noise weighting curves

- Fixed center frequency, increasing noise bandwidth
  - Spike count increases above auditory bandwidth (dotted line)
Results — Comparison to Other Measures

- Comparison to widely used acoustic indicators
  - Noise level in decibels with “A” frequency weighting, denoted “dB(A)”
  - Loudness (a psychoacoustic model of perceived sound intensity)

→ Significantly lower prediction error ($p < 0.01$) on 2 datasets

<table>
<thead>
<tr>
<th>Measure</th>
<th>Prediction Error (lower values are better)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Set 1</td>
</tr>
<tr>
<td>Weighted Level [dB(A) SPL]</td>
<td>0.230</td>
</tr>
<tr>
<td>Mean Loudness [sone]</td>
<td>0.257</td>
</tr>
<tr>
<td>5th Percentile Loudness [sone]</td>
<td>0.191</td>
</tr>
<tr>
<td>5th Percentile Density [spikes/s]</td>
<td>0.087**</td>
</tr>
</tbody>
</table>
Results — (In)sensitivity to Parameters

- Robust to changes in dictionary design

![Bar chart showing Stacked Prediction Errors rmse^3rd for different sets and number of kernels in dictionary.](chart)

- Number and Type of Kernels in Dictionary:
  - Set 1
  - Set 2
  - Set 3
Conclusion

• We are doing audio processing, not speech processing

• Number of “spikes” reflects the level and type of noise

• Sparsity of noise over time highly correlates with perceived intrusiveness

• Efficient coding hypothesis offers a different interpretation of intrusiveness:
  • Complexity of the input stream to the auditory system
  • Activations of nerve spike populations in response to noise
Thank You for Your Attention

Thanks to
• Laboratory of Electromagnetics and Acoustics (LEMA), EPFL
• SwissQual AG
• Dr. Marc Ferras and Dr. Mathew Magimai-Doss for useful discussions