Learning Mid-Level Codes for Natural Sounds

Wiktor Młynarski, Josh H. McDermott
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

mlynar@mit.edu

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216

CENTER FOR
Brains
Minds+
Machines

Motivation and background

In the visual domain, models and physiological studies suggest mid-level representations of natural stimuli.

Little is known about comparable auditory structures of intermediate complexity. Are there auditory analogues of contours, curvature and shapes?

We propose a probabilistic model of natural sounds in an attempt to answer the following questions:

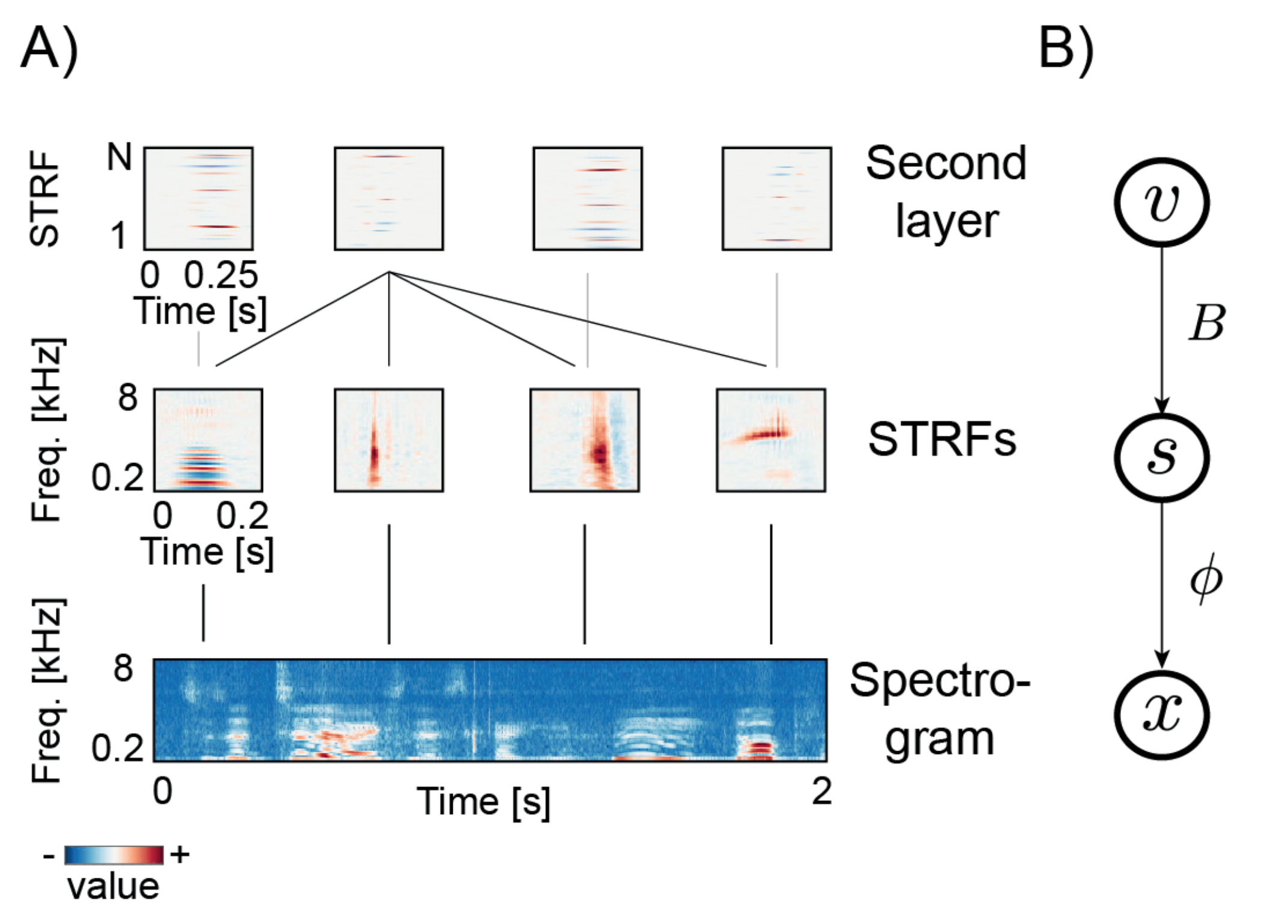
1) What are the naturally occuring combinations of basic spectrotemporal features (STRFs)?

2) How can more abstract representations be developed from linear STRFs?

3) What can be learned about mid-level audition from natural patterns of STRF combinations?

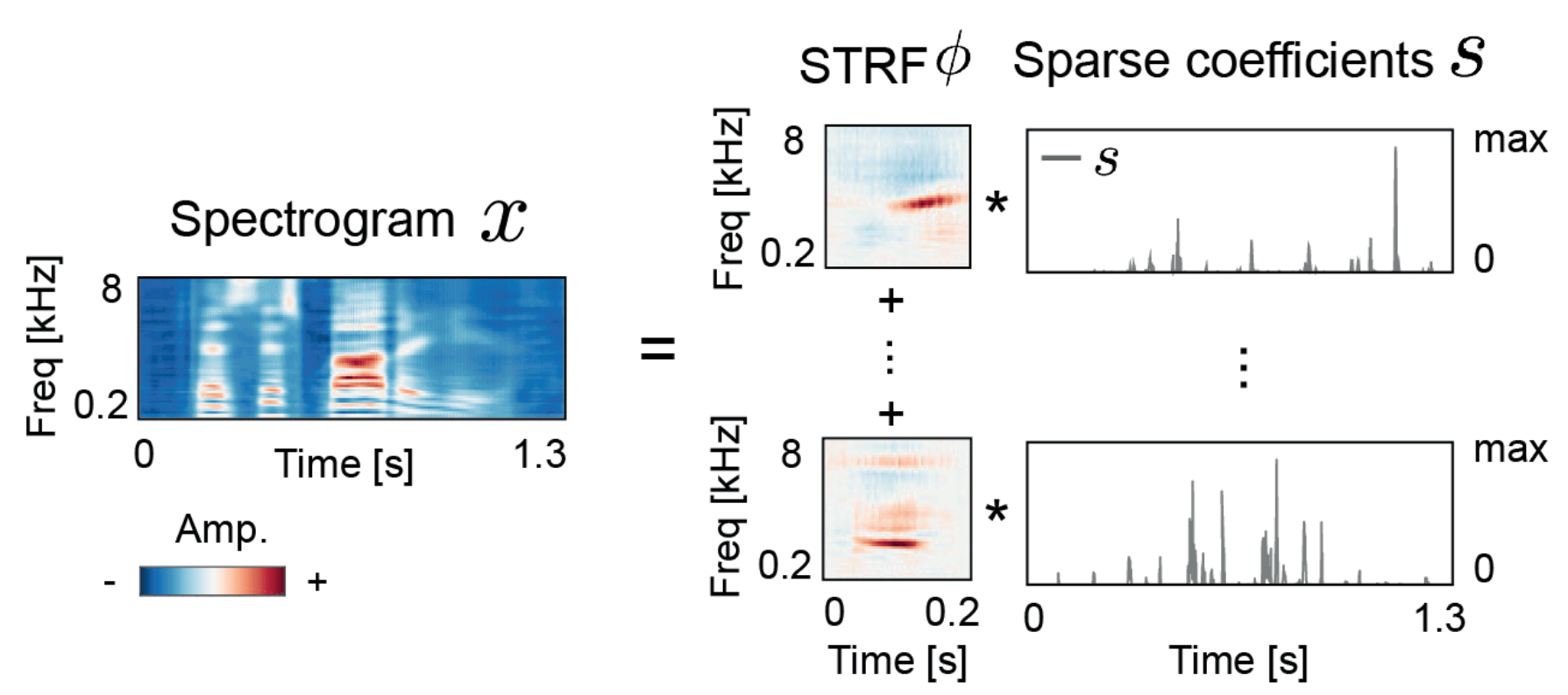
Model overview

Two layer hierarchical generative model.
First layer: decomposes spectrogram into linear composition of spectrotemporal features (STRFs) convolved with sparse, non-negative activation time courses.
Second layer: encodes STRF coactivation patterns.



Overview of the hierarchical model A) A spectrogram (first-row) is encoded by a set of spectrotemporal features (second row).
The temporal patterns of multiple STRF activations are encoded by second-layer features. B) A graphical model depicting statistical dependencies among variables.

First layer - sparse spectrogram features



$$p(s_{i,t}|\lambda_{i,t}) = Exp(\lambda_{i,t})$$

$$x_{t,f} = \left[\sum_{i}^{N} \phi_{i,f} * s_i\right]_t + \eta$$

$$\eta \sim \mathcal{N}(0, \sigma^2)$$

First layer scheme. A spectrogram of arbitrary length is represented as a sum of spectrotemporal features convolved with time-courses of corresponding coefficients. Coefficients are non-negative and have sparse distributions (i.e. remain close to zero most of the time)

Diversity of learned spectrogram features

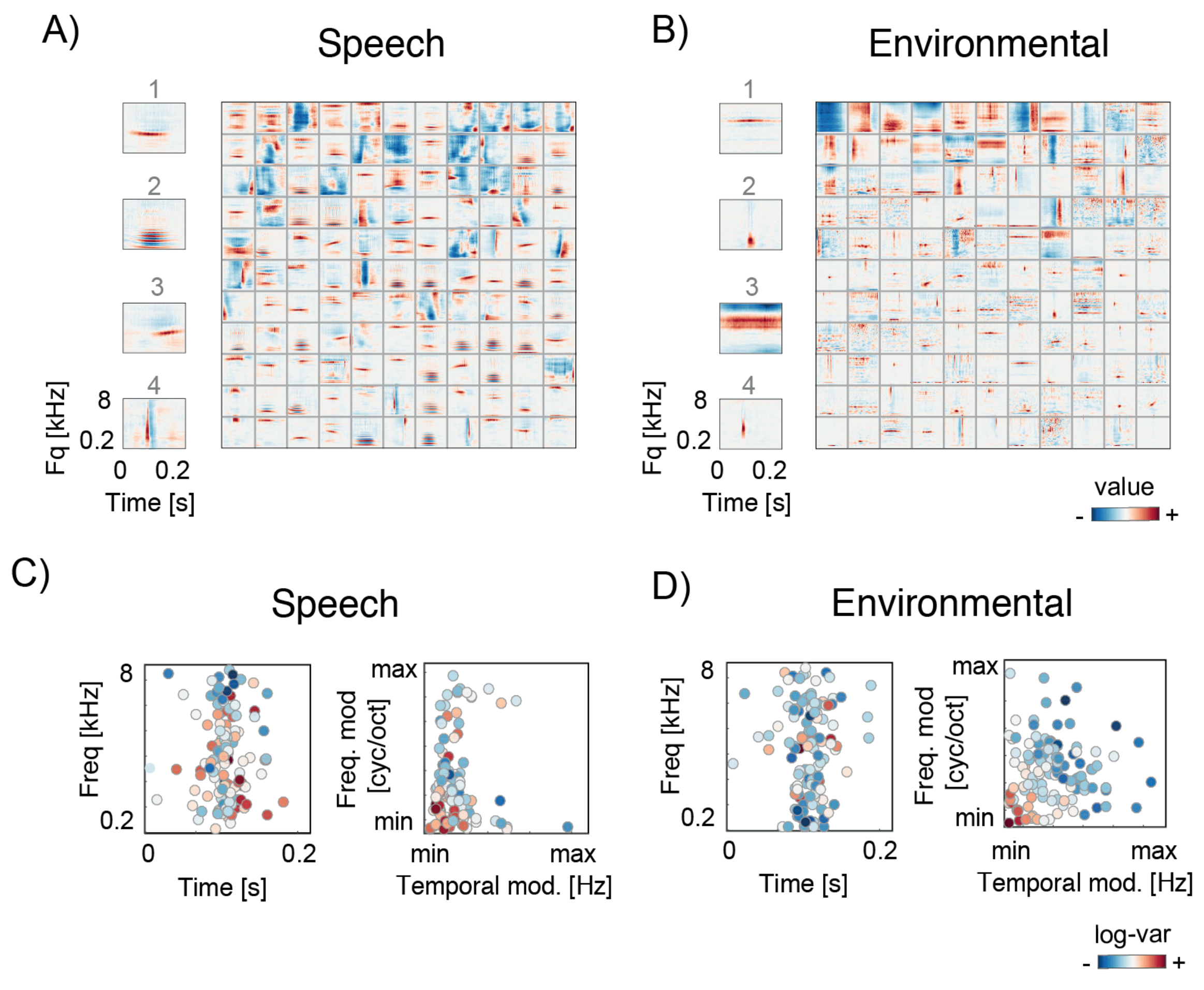
Spectrotemporal features learned by the first layer.

A) Population of STRFs learned from a speech database.

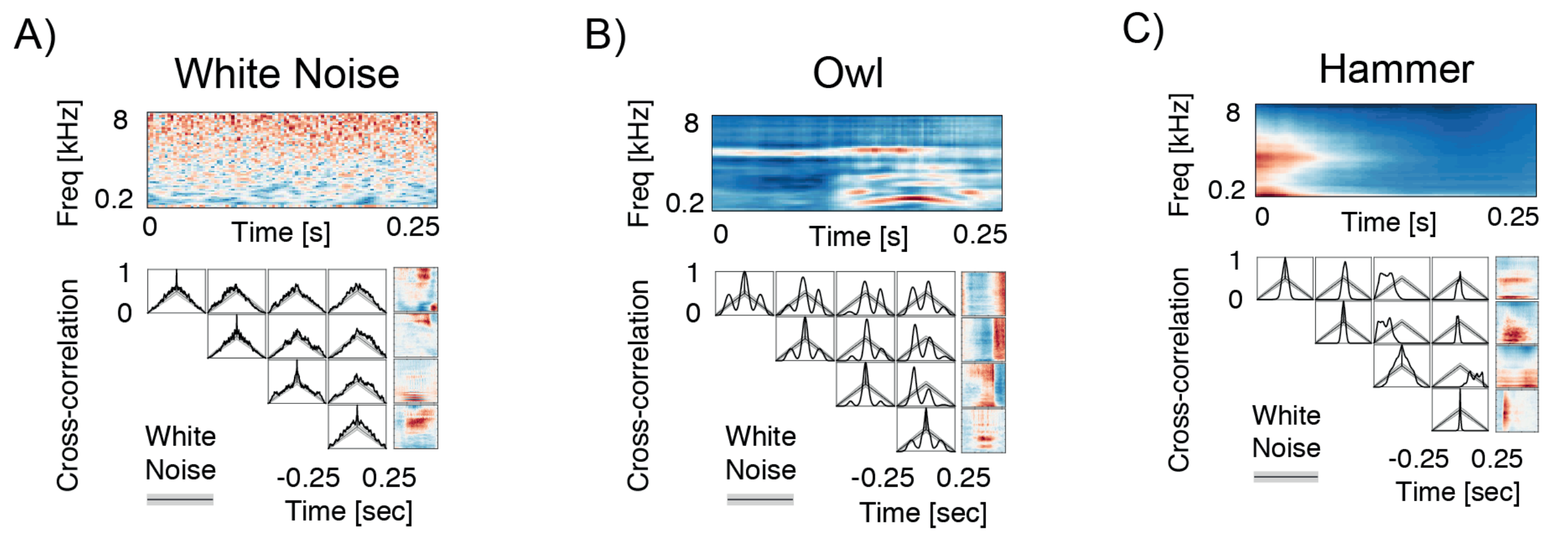
B) Population of STRFs learned from a database of environmental sounds.

Representative STRFs for each corpus were magnified and are numbered from 1-4. They include clicks, single harmonics and harmonic stacks.

C, D) Representations of STRF populations on time-frequency (left) and spectral-temporal modulation (right) planes. Each dot corresponds to a single STRF, and its color encodes log-magnitude averaged over entire training dataset.



Signatures of mid-level auditory structure



Dependencies among STRF coefficients.

Locally, STRF activations reveal strong correlations when encoding natural sounds such as an owl sound (B) and a hammer hit (C). This is not the case for white noise (A)

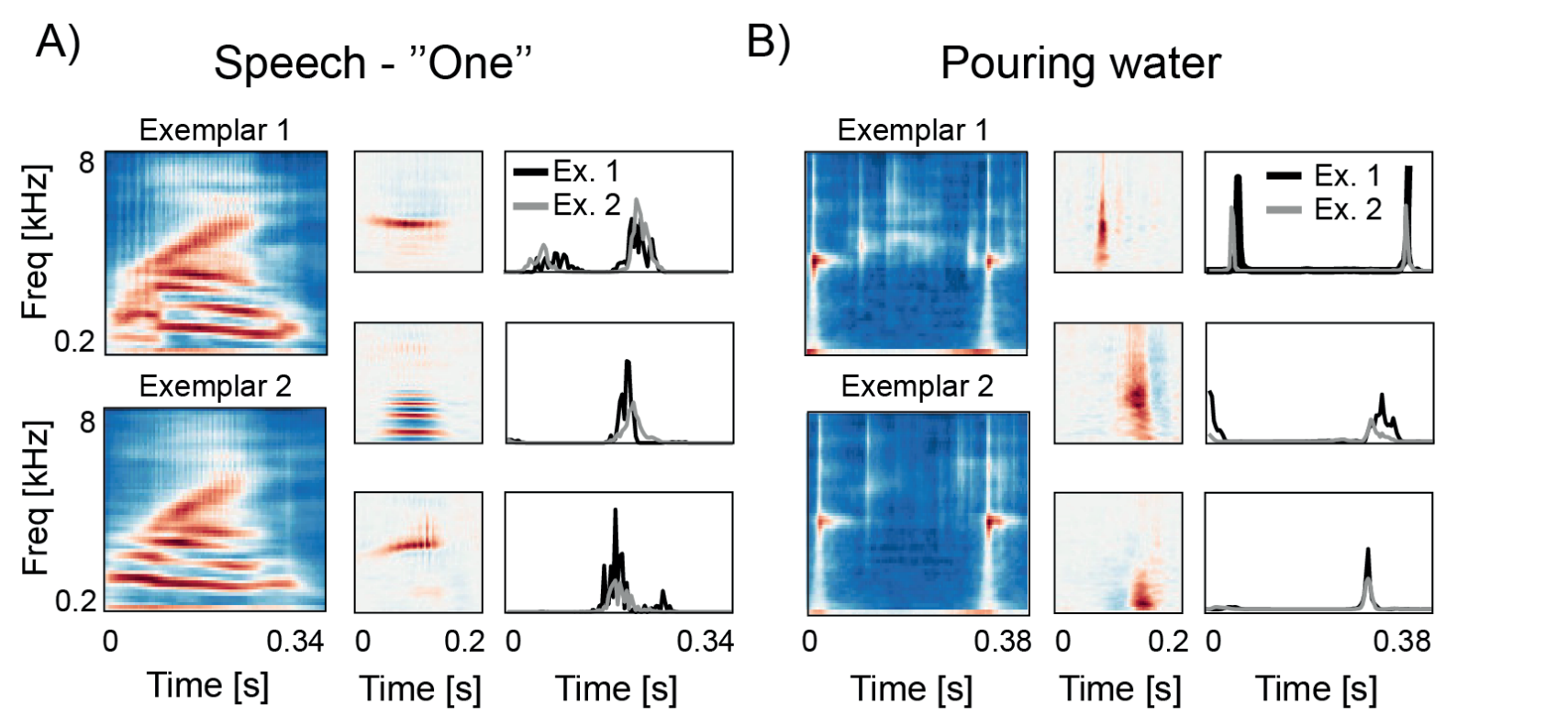
Strong STRF correlations imply presence of higher-order auditory structures.

Second layer - STRF combinations - mid-level auditory code

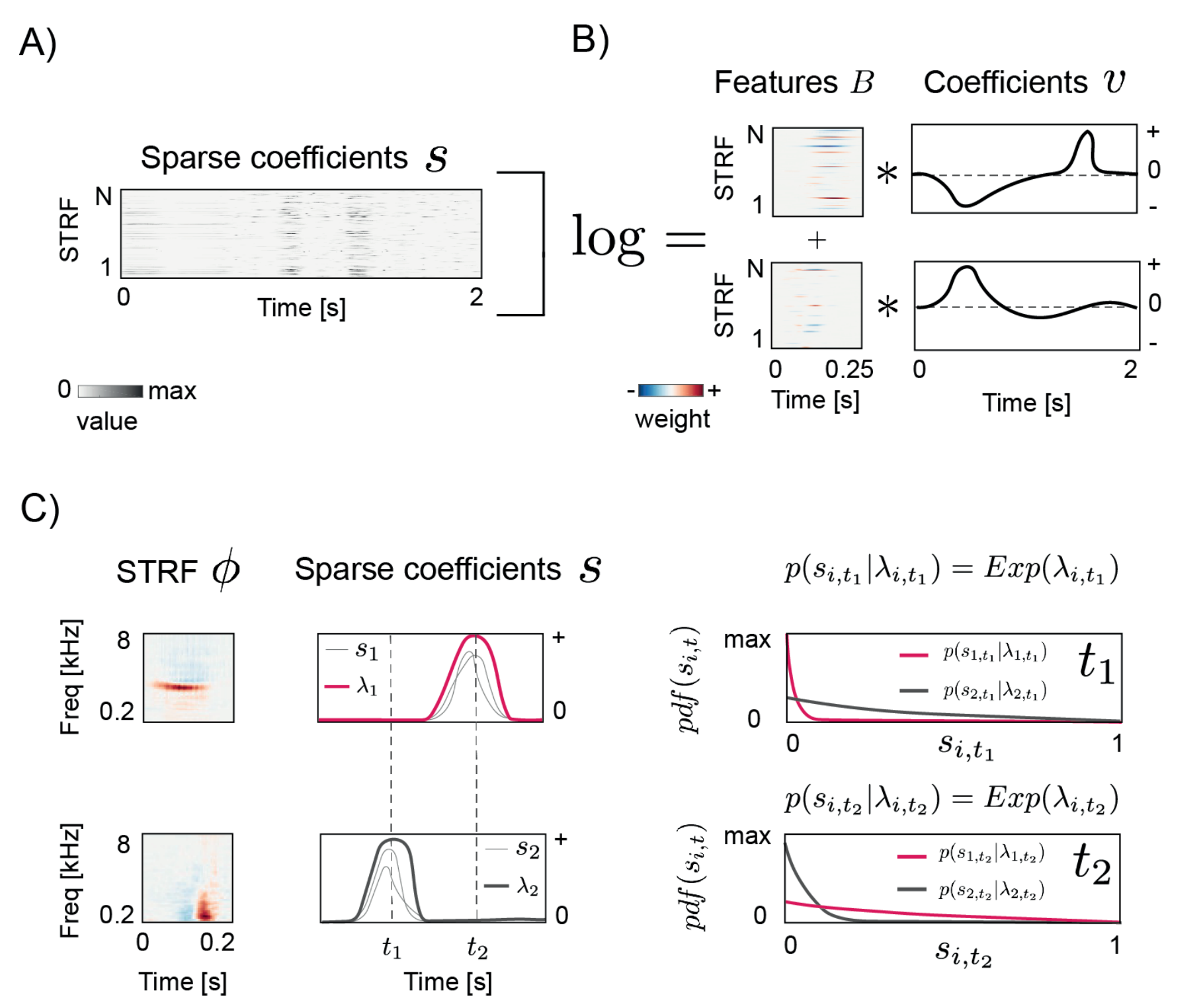
Modelling the magnitude of STRF activations.

Coefficient trajectories of STRFs are depicted for two exemplars of different natural sounds.

While minor difference between coefficient trajectories (gray and black lines) are visible, they preserve the same global structure which is inherent to each sound.



Encoding time-varying magnitudes of STRF activations



$$p(v_{j,t}) \propto \exp\Big(-\alpha|v_{j,t}|\Big)$$

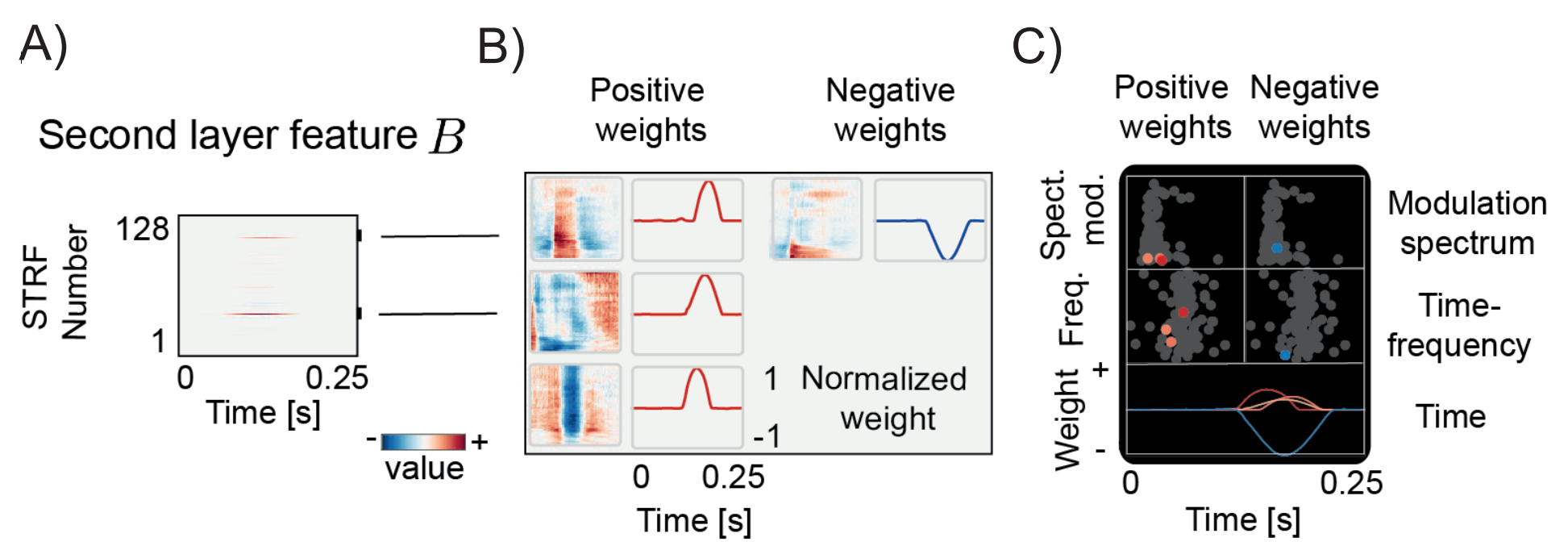
$$\lambda_{i,t} = \exp\left[\sum_{j=1}^{M} B_{j,i} * v_j + \rho_i\right]_t$$

$$p(s_{i,t}) = Exp(\lambda_{i,t}) = \frac{1}{\lambda_{i,t}} \exp\left[-\frac{s_{i,t}}{\lambda_{i,t}}\right]$$

Second layer scheme A), B) An array of STRF activations serves as an input to the second layer. It uses a population of features B to encode the logarithm of STRF activation magnitudes

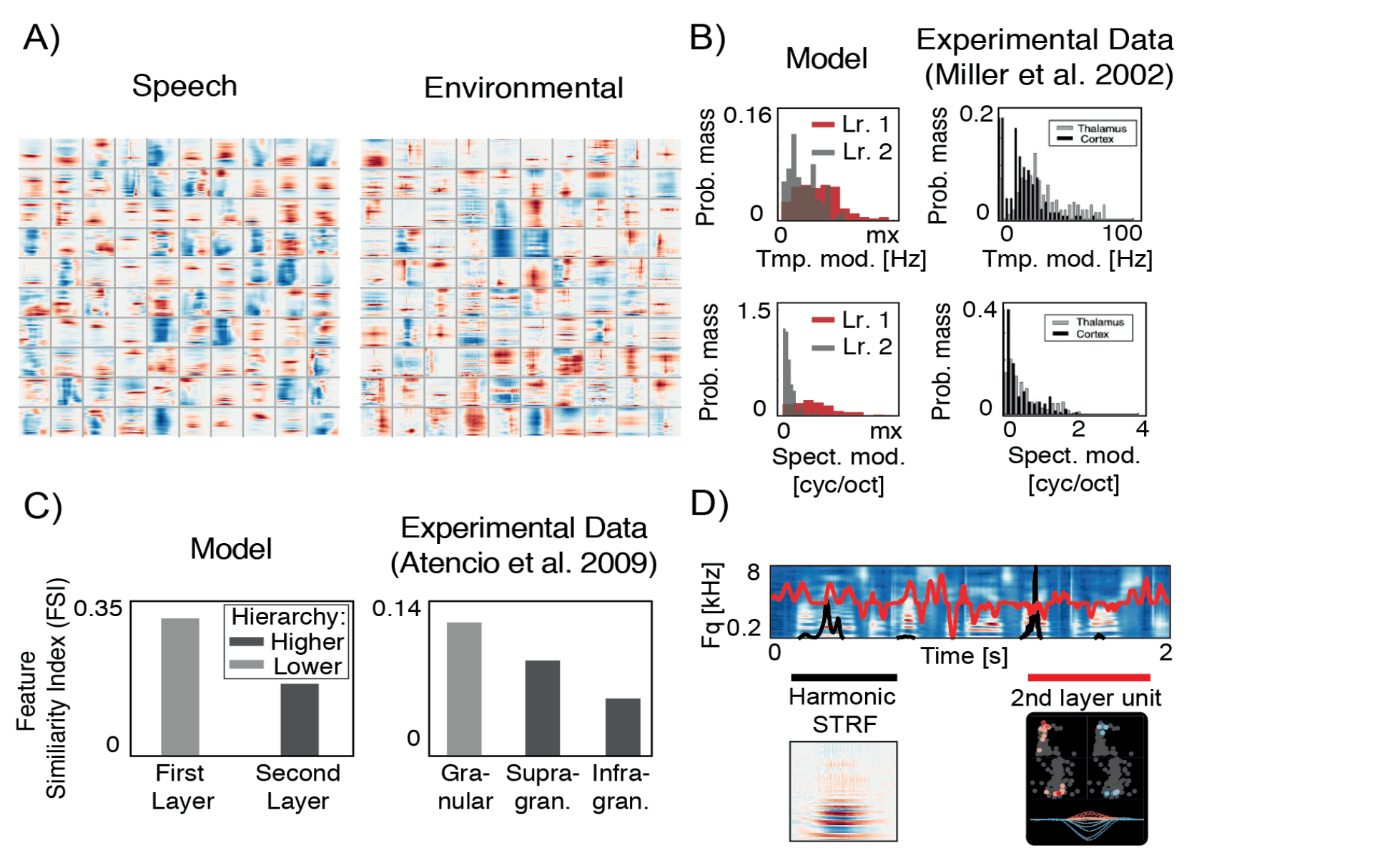
C) Ilustration of time-changing distributions of STRF activations.

Visualizing second layer units



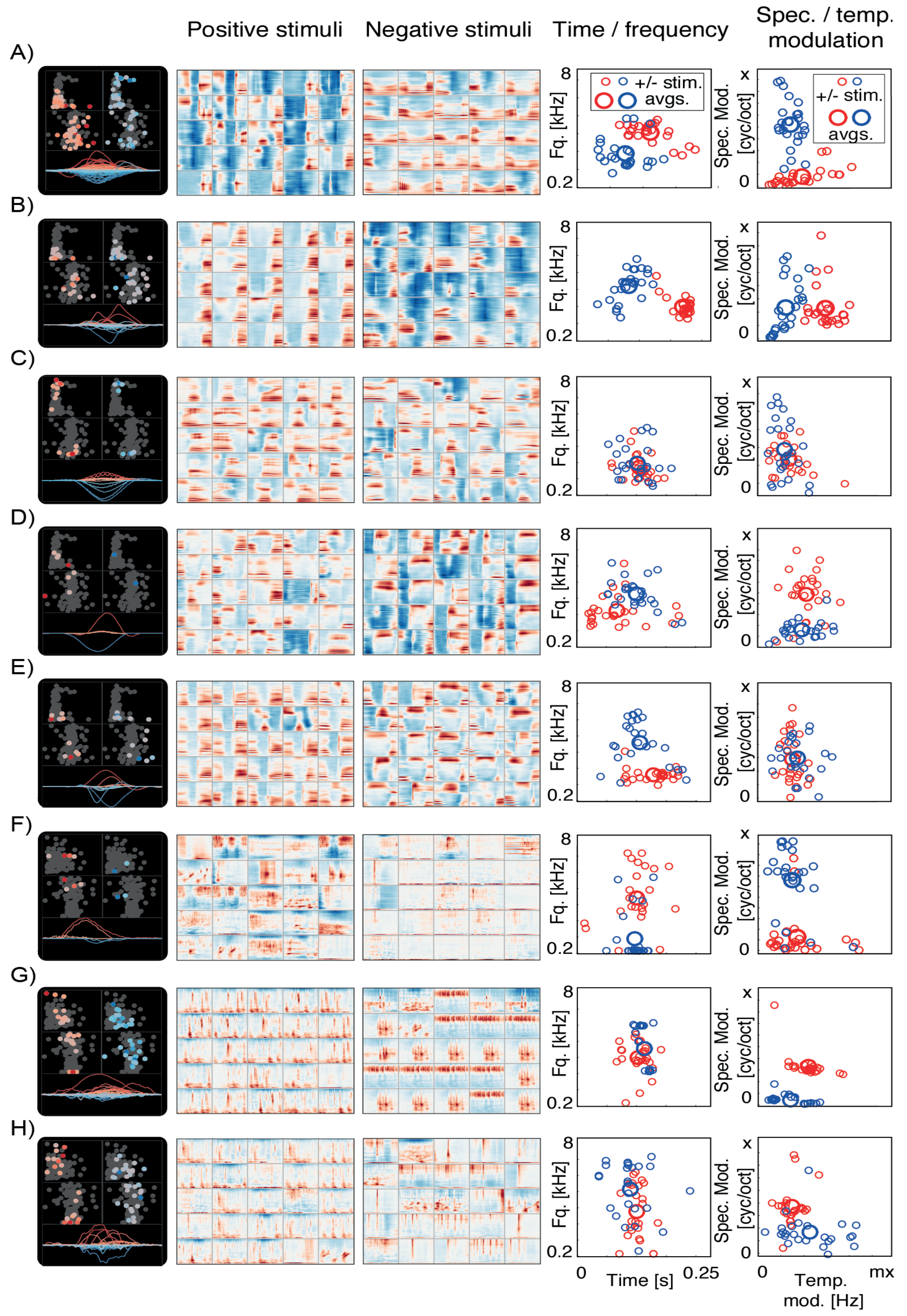
A) A second layer feature encoding temporal activations of STRFs. B) STRFs with weights significantly deviating from 0. C) STRFs with positive weights are plotted as red dots and negative weighted STRFs as blue dots

Comparison to neural data



A) Receptive fields estimated for second layer units.
B) Modulation spectra of model receptive fields and data from auditory thalamus and cortex [1]. C) Comparison of Feature Selectivity Index computed for model layers and layers of the auditory cortex [2]. D) The example activation of the first layer unit is locked to the presence of a preferred stimulus. Responses of the second layer unit are visibly less specific.

"Excitation-Inhibition" patterns in second-layer units



"Excitatory" and "inhibitory" stimuli

Each row depicts a second-layer feature in the leftmost column.

Second and third columns:
25 stimuli eliciting strong positive and negative responses of the corresponding unit.
Fourth and fifth columns: positive and negative stimuli visualized as red and blue circles, respectively, in time-frequency and modulation planes. Large circles correspond to centroids of positive and negative stimulus clusters.

Second-layer units can be grouped into four classes, based on the separation of the positive and negative stimuli in the time-frequency plane and/or modulation plane.

First class: stimuli are separated in both domains (Units A, B and F)

Second class: stimuli separated by spectrotemporal modulation (Units D, G and H).

Third class: stimuli separated only in frequency plane (Unit E).

Fourth class: No visible separation in either domain (Unit C).

Literature:
[1] Miller, Lee M., et al. "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex." J of Neurophys, 2002,
[2] Atencio, Craig A. et al. "Hierarchical computation in the canonical auditory cortical circuit." PNAS, 2009