Learning spectro-temporal representation of complex sounds with parameterized neural networks

Rachid Riad, Julien Karadayi, Anne-Catherine Bachoud-Levi and Emmanuel Dupoux

https://arxiv.org/abs/2103.07125

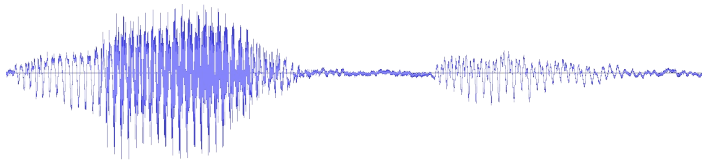
Examples of auditory inputs

Examples of auditory inputs

1. People chatting inside 
2. Car alarm 
3. Bird 

Examples of auditory inputs

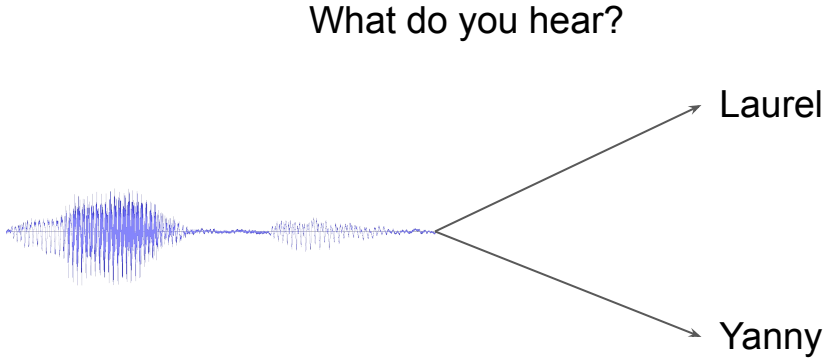
1. People chatting inside
2. Car alarm
3. Bird



Audition

How do we obtain meaningful information from sounds?

Psychology

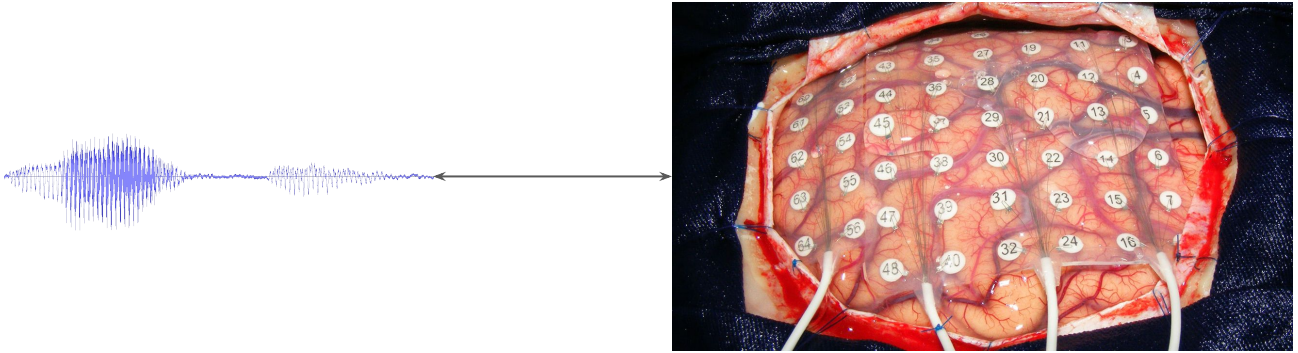


Audition

How do we obtain meaningful information from sounds?

Auditory Neuroscience

How sounds are represented in the brain?

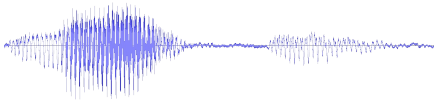


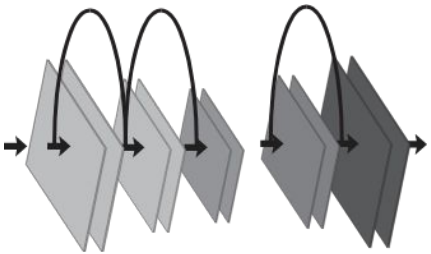
Audition

How do we obtain meaningful information from sounds?

Machine Listening

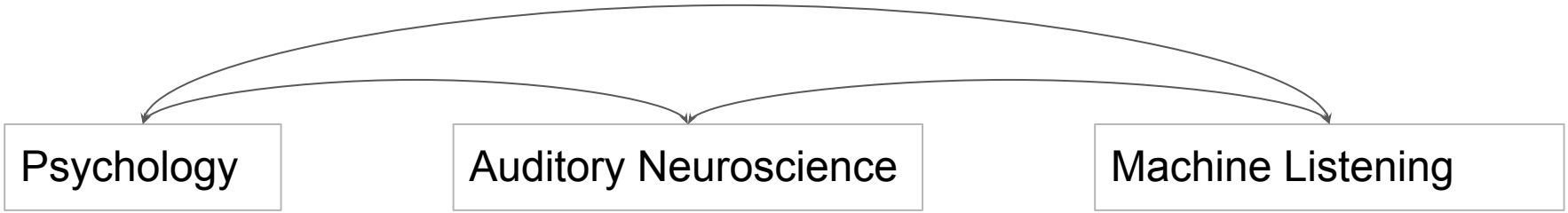
Can we design an algorithm to deal with sounds?





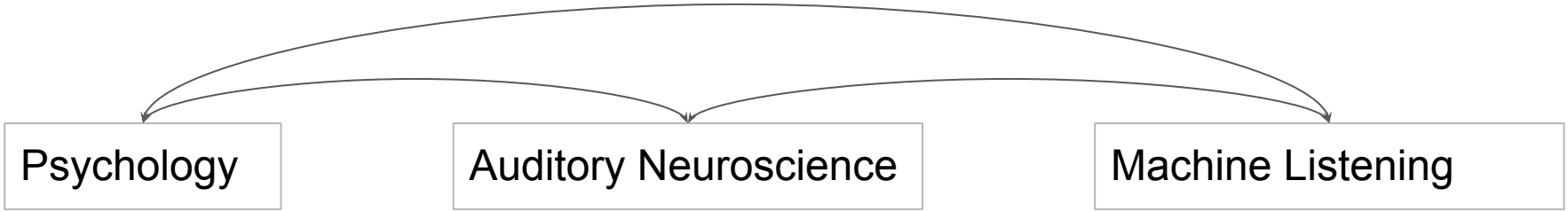
Audition

How do we obtain meaningful information from sounds?



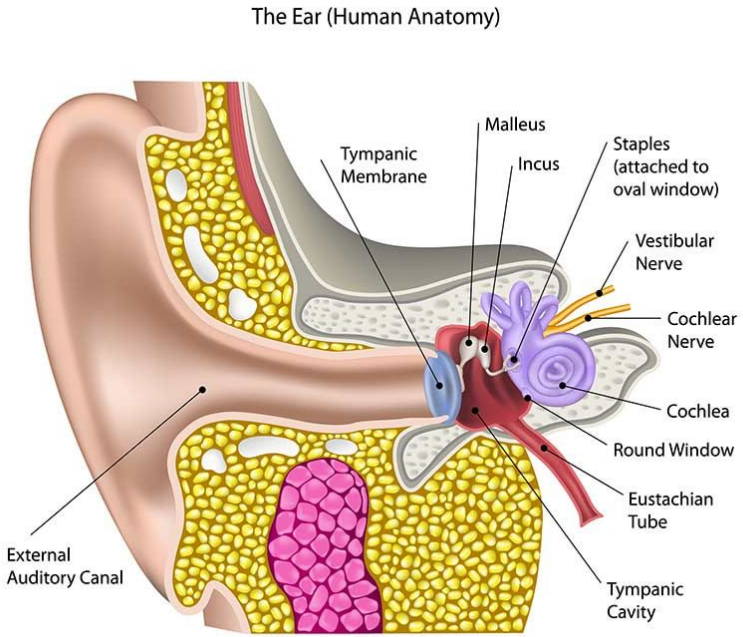
Audition

How do we obtain meaningful information from sounds?

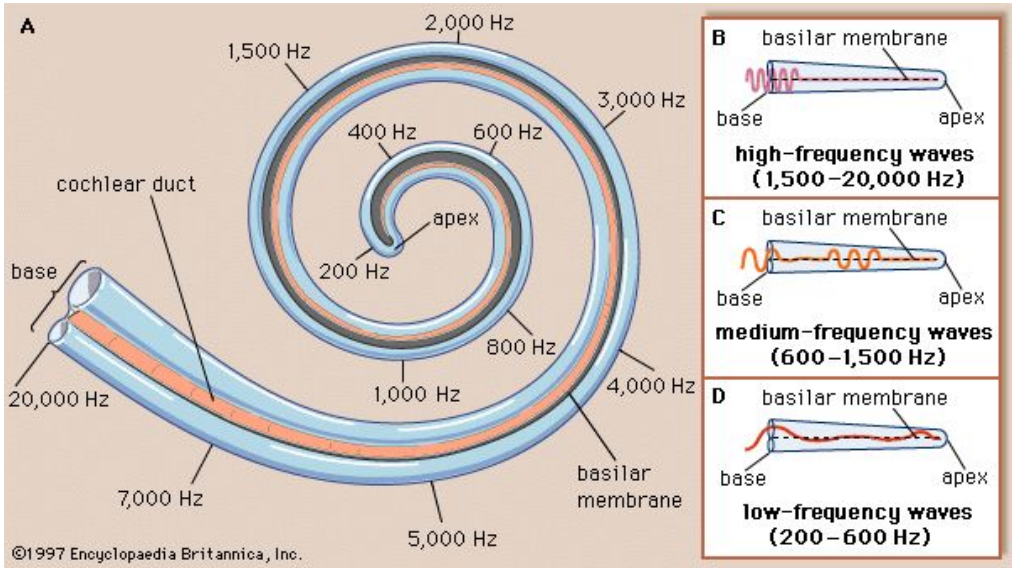


Not so much in practice!!

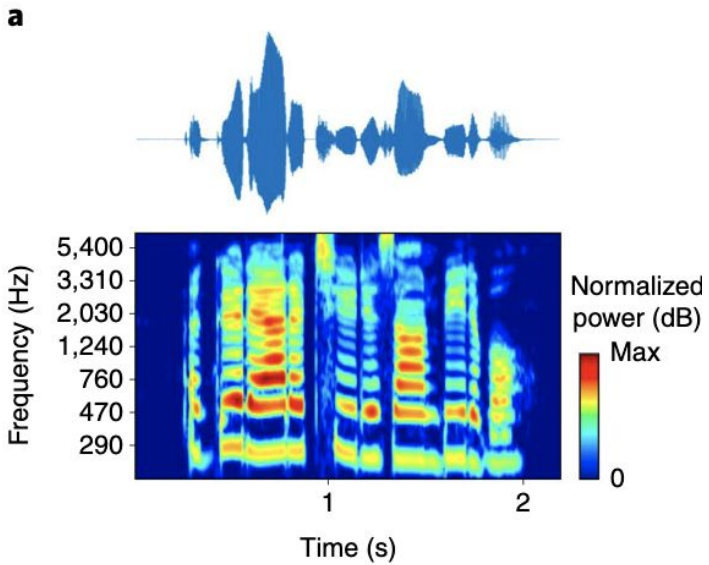
Classic auditory models



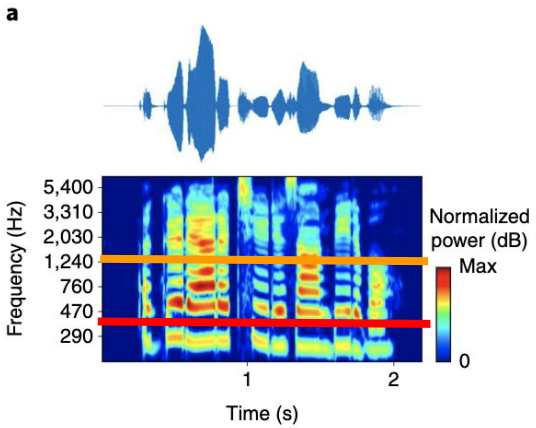
Classic auditory models

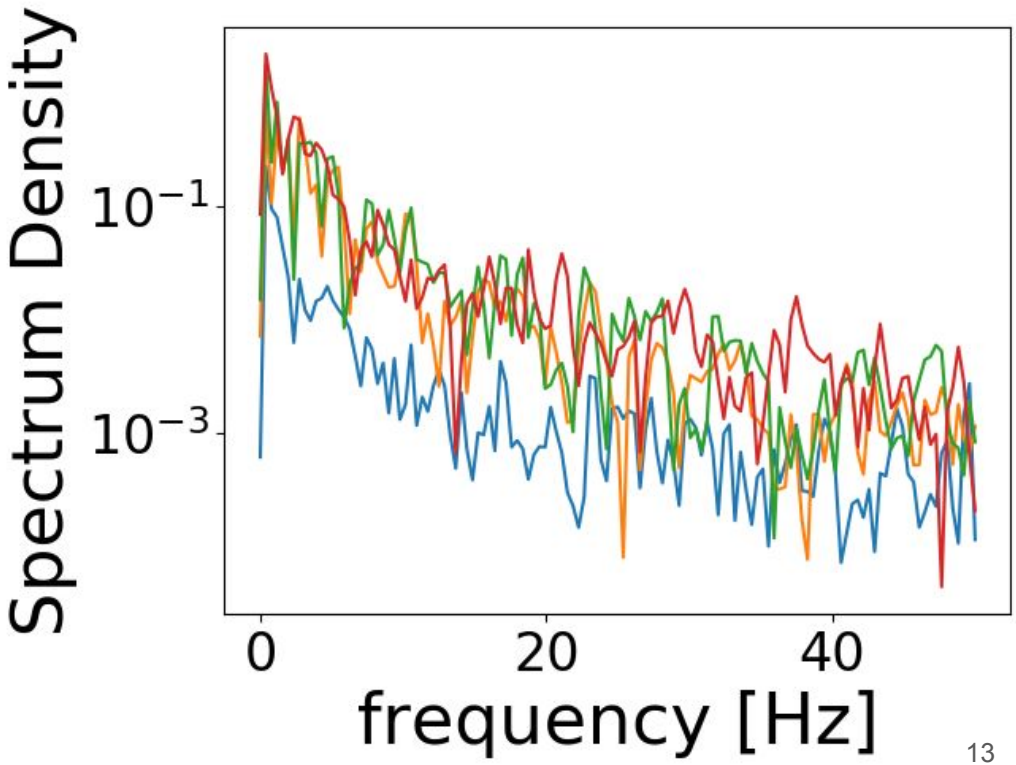


Classic auditory models

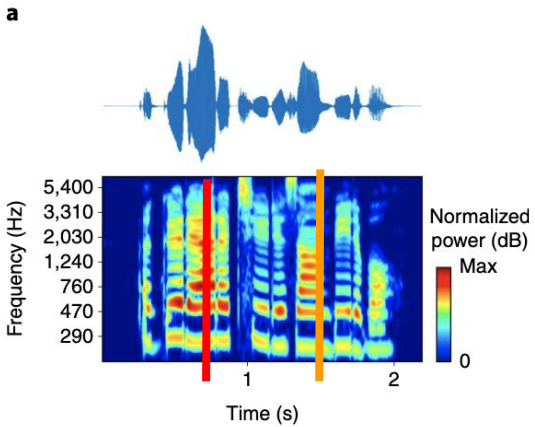


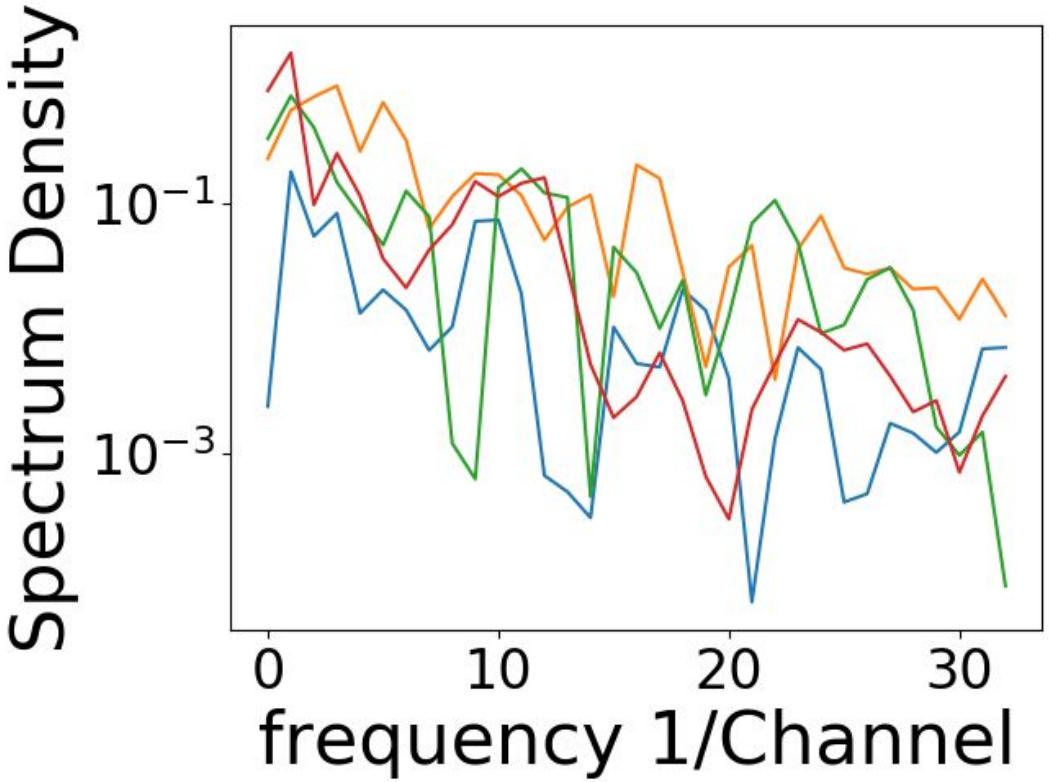
Modulations of the spectrogram



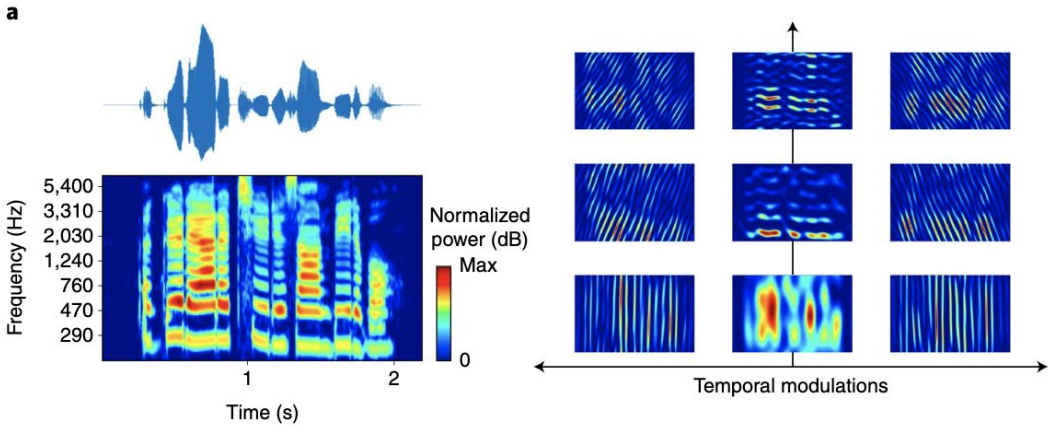


Modulations of the spectrogram

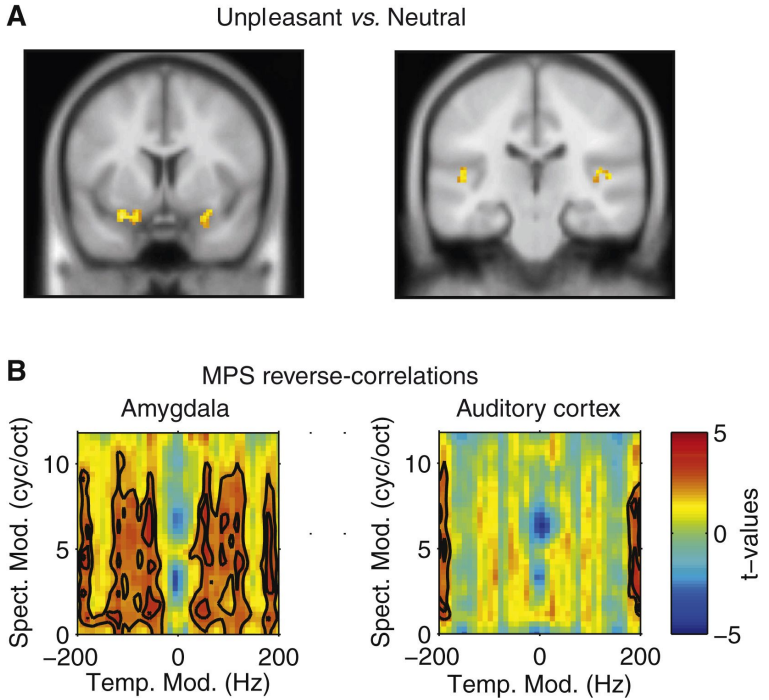




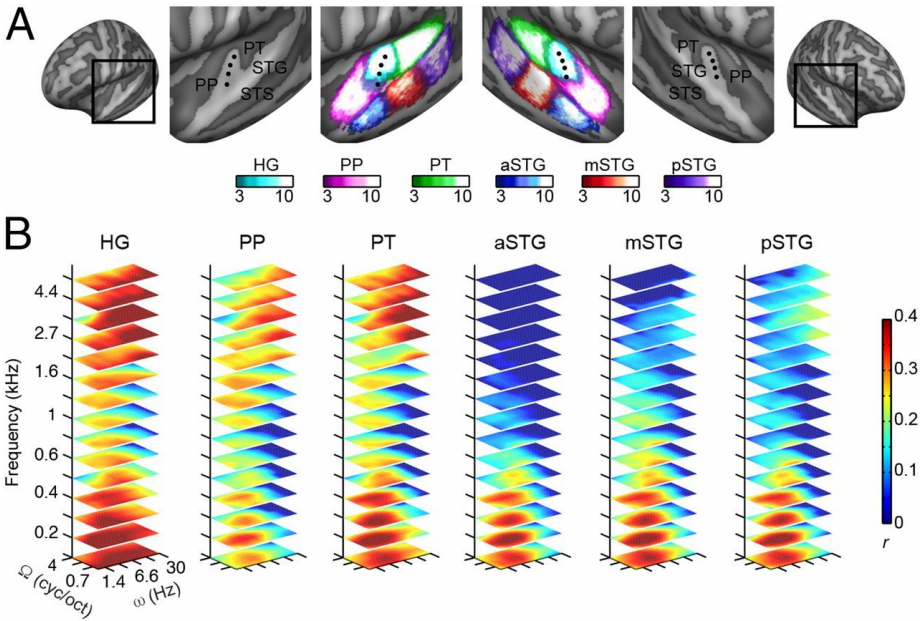
Classic auditory models



Classic auditory models

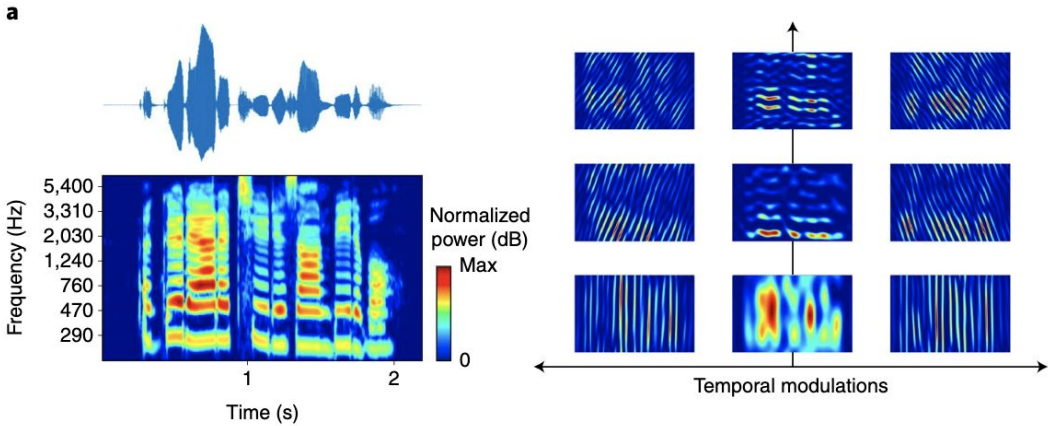


Arnal et al. 2015



Santoro et al. 2017

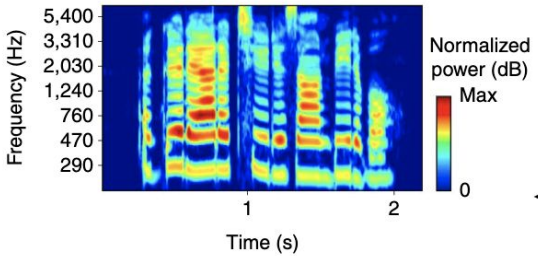
Classic auditory models

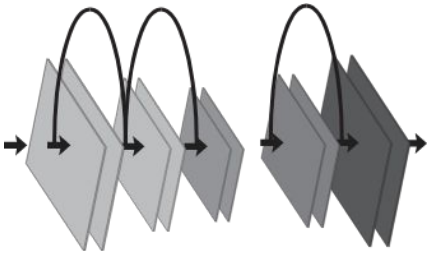


- Does not account for **behaviour**
- Does not correlate much with the **brain**

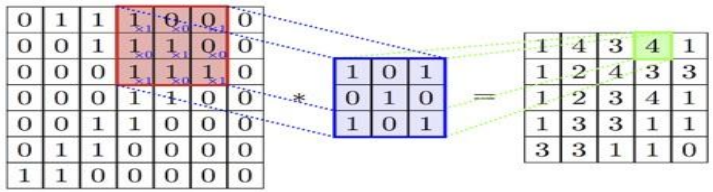
Deep Learning models as models of audition

Human level performance reached by Deep Neural Networks based on the spectrogram

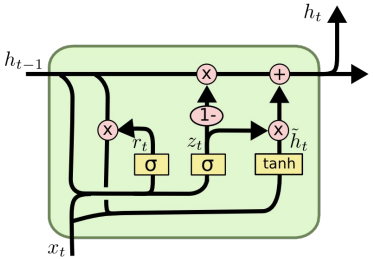




Convolution layers



Recurrent layers



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

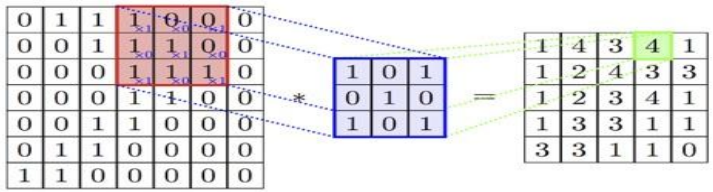
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Deep Learning models as models of audition

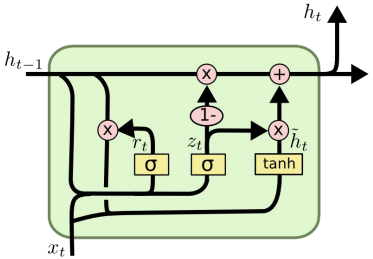
Human level performance reached by Deep Neural Networks based on the spectrogram

Not **interpretable** for neuroscientists and psychologists

Convolution layers



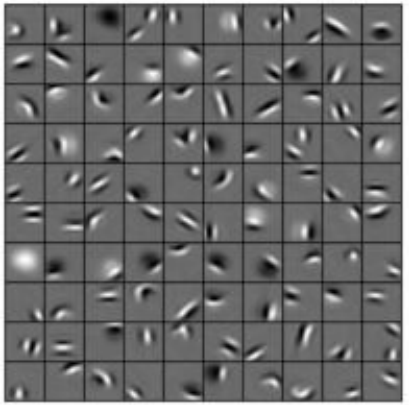
Recurrent layers



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

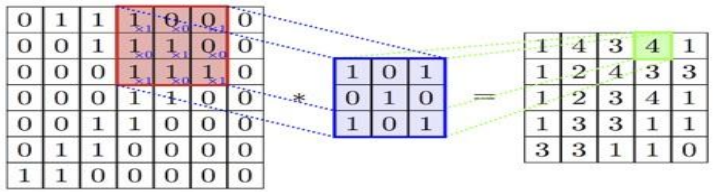
Deep Learning models as models of audition

Human level performance reached by Deep Neural Networks based on the spectrogram

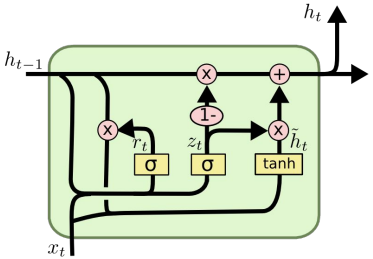


Not **interpretable** for neuroscientists and psychologists

Convolution layers

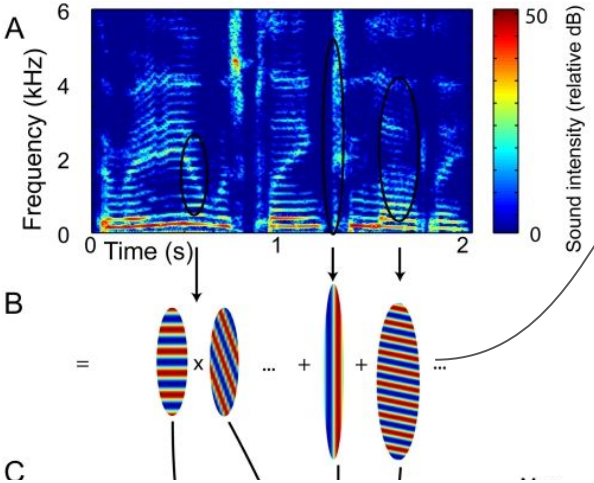


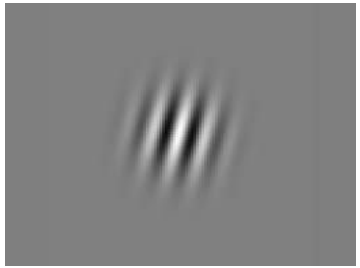
Recurrent layers



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

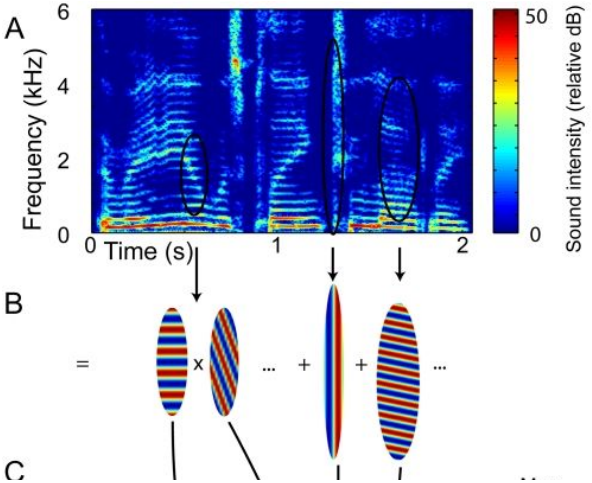
Modulations of the spectrogram

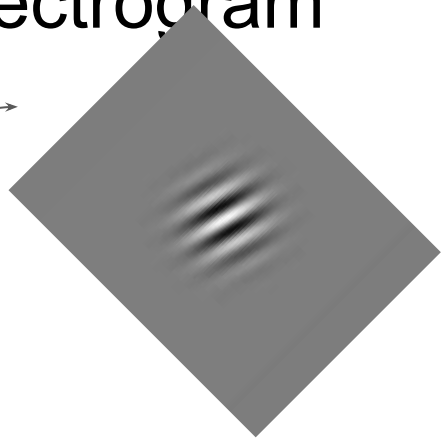




Like the Gabor patterns found in Vision processing like in V1

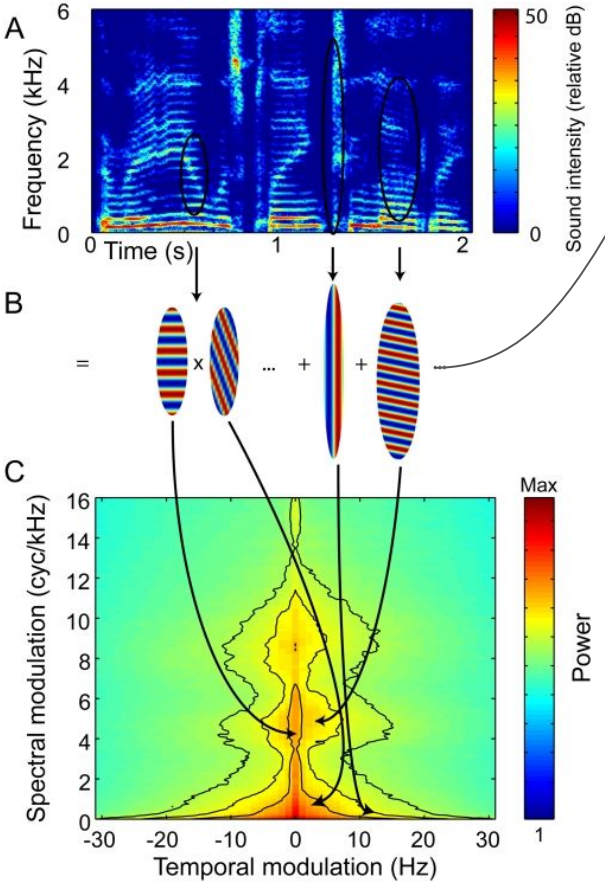
Modulations of the spectrogram

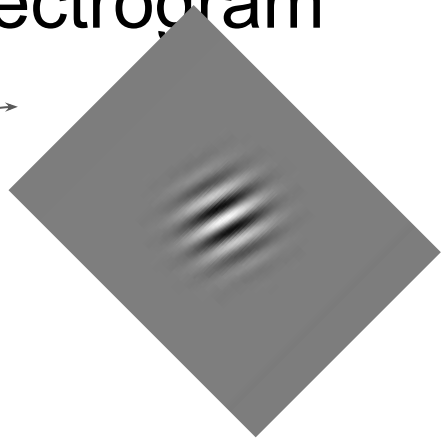




Like the Gabor patterns found in Vision processing like in V1

Modulations of the spectrogram



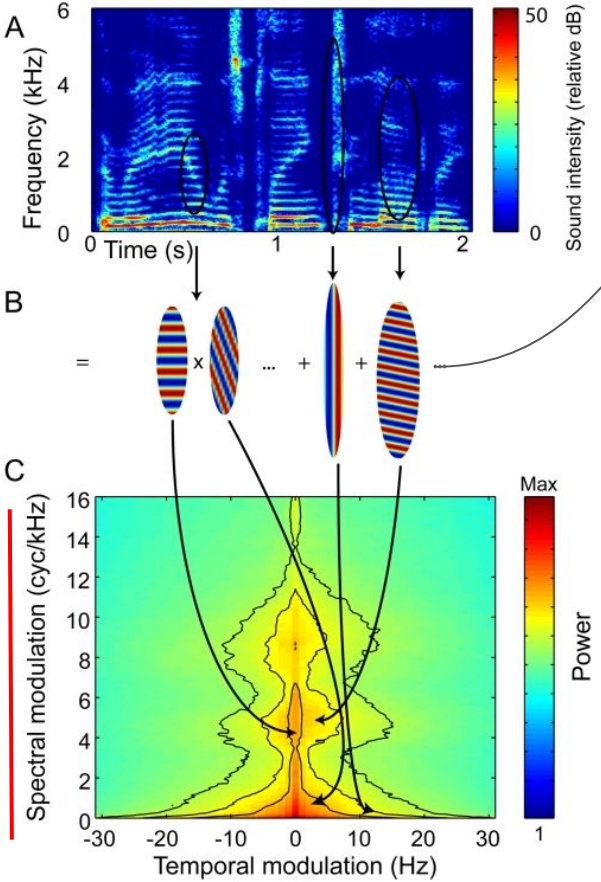


Like the Gabor patterns found in Vision processing like in V1

≠

Distribution is different in auditory domain

Modulations of the spectrogram



$$g_k(t,f) = s_k(t,f)w_k(t,f)$$

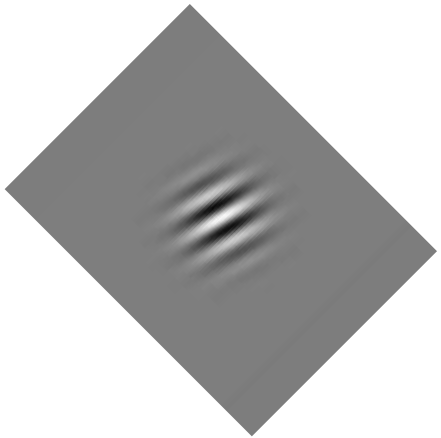
$$s_k(t,f) = \frac{1}{2\pi\sigma_t\sigma_f}e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)}$$

$$w_k(t,f) = e^{j(2\pi(\omega_k t+\Omega_k f))}$$

$$w_k(t,f) = e^{j(2\pi(F_k R_{\gamma_k}))}$$

$$R_{\gamma_k} = t\cos(\gamma_k) + f\sin(\gamma_k)$$

How to incorporate modulations in deep models?



$$g_k(t,f) = s_k(t,f)w_k(t,f)$$

$$s_k(t,f) = \frac{1}{2\pi\sigma_t\sigma_f} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)}$$

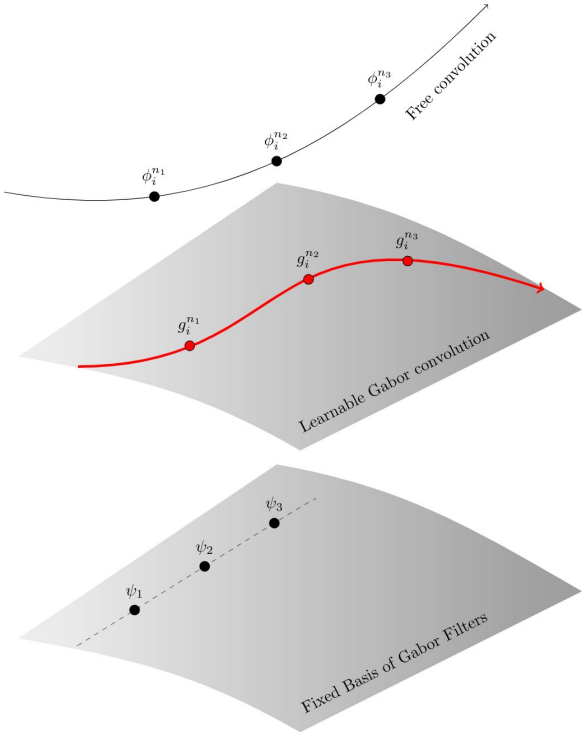
$$w_k(t,f) = e^{j(2\pi(\omega_k t+\Omega_k f))}$$

$$w_k(t,f) = e^{j\left(2\pi\left(F_k R_{\gamma_k}\right)\right)}$$

$$R_{\gamma_k} = t\cos(\gamma_k) + f\sin(\gamma_k)$$

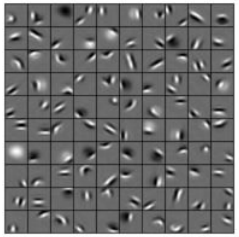
How to incorporate modulation in deep models?

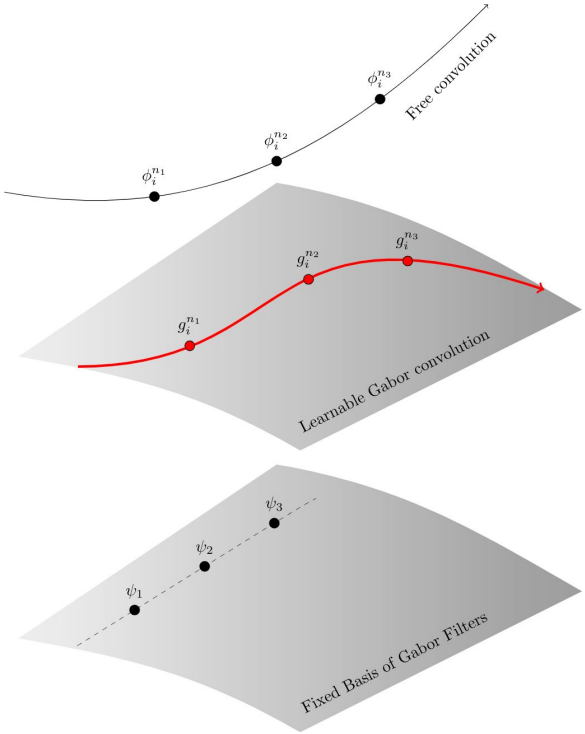
The idea: Stay in the gabor Domain for the convolution, but find the relevant ones

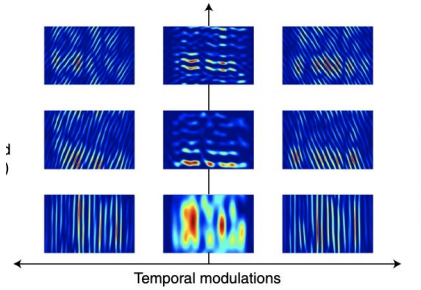


How to incorporate modulation in deep models?

The idea: Stay in the gabor Domain for the convolution, but find the relevant ones

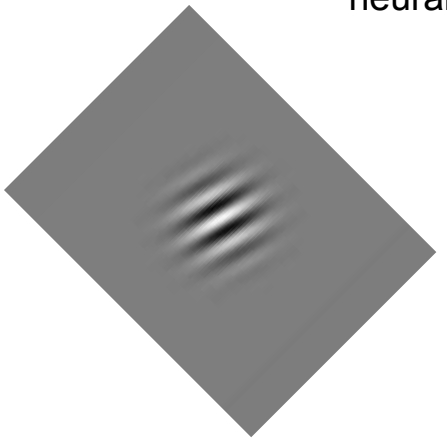






How to incorporate modulation in deep models?

How to implement the idea: Integrate convolutions into a neural network and backprop only a subset of parameters



$$g_k(t,f) = s_k(t,f)w_k(t,f)$$

$$s_k(t,f) = \frac{1}{2\pi\sigma_t\sigma_f} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)}$$

$$w_k(t,f) = e^{j(2\pi(\omega_k t+\Omega_k f))}$$

$$w_k(t,f) = e^{j(2\pi(F_k R_{\gamma_k}))}$$

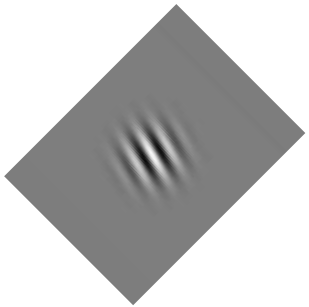
$$R_{\gamma_k} = t\cos(\gamma_k) + f\sin(\gamma_k)$$

Learn only $(\sigma_t, \sigma_f, R_{\gamma_k}, F_k)$

4 params instead of 111*9

How to incorporate modulation in deep models?

How to implement the idea: Integrate convolutions into a neural network and backprop only a subset of parameters

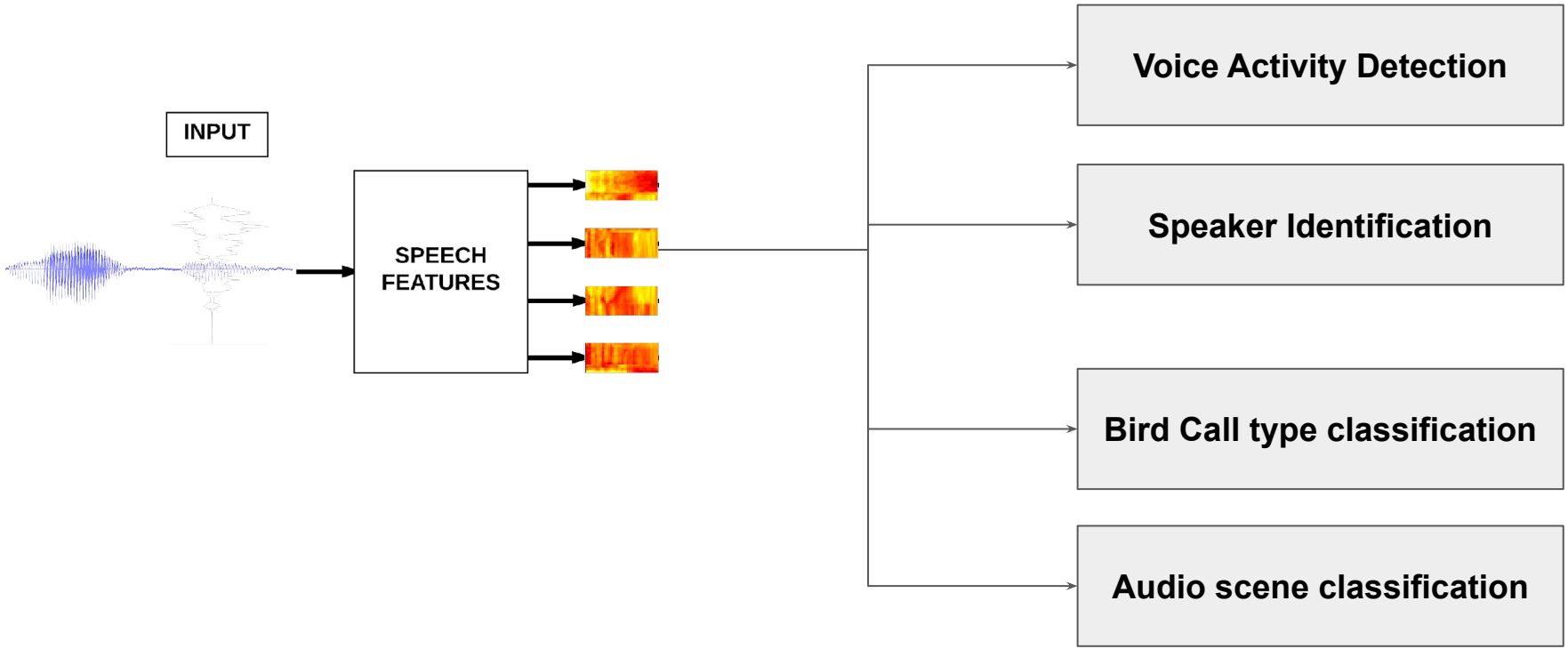


$$
\begin{aligned}
g_k(t,f) &= s_k(t,f)w_k(t,f) \\
s_k(t,f) &= \frac{1}{2\pi\sigma_t\sigma_f} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)} \\
w_k(t,f) &= e^{j(2\pi(\omega_k t+\Omega_k f))} \\
w_k(t,f) &= e^{j(2\pi(F_k R_{\gamma_k}))} \\
R_{\gamma_k} &= t\cos(\gamma_k) + f\sin(\gamma_k)
\end{aligned}
$$

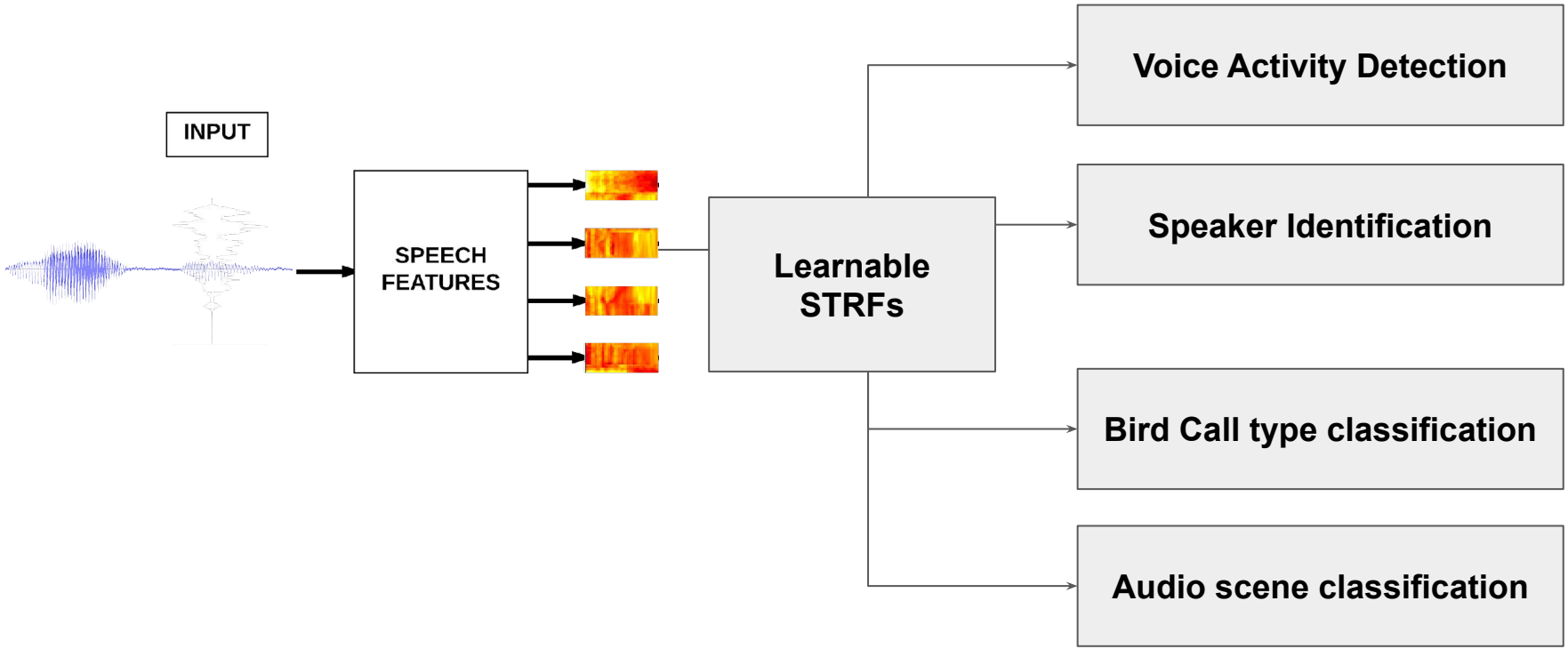
Learn only $(\sigma_t, \sigma_f, R_{\gamma_k}, F_k)$

4 params instead of 111*9

Our pipeline for audio processing



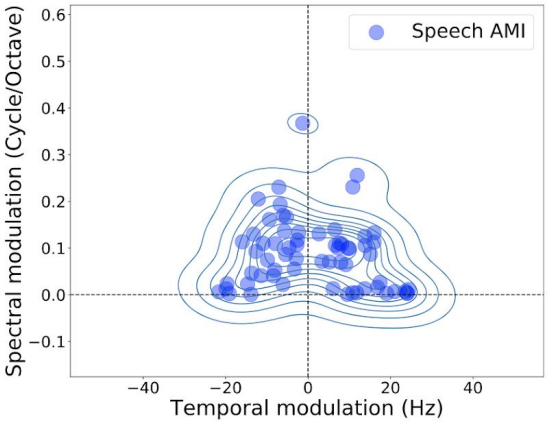
Our pipeline for audio processing

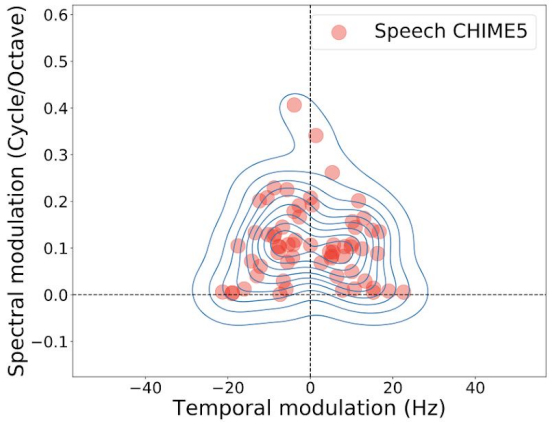


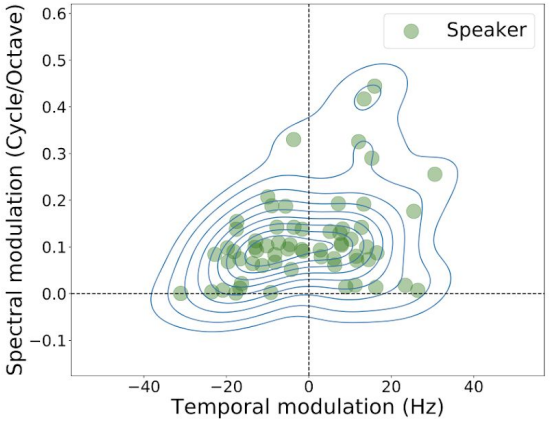
Engineering results

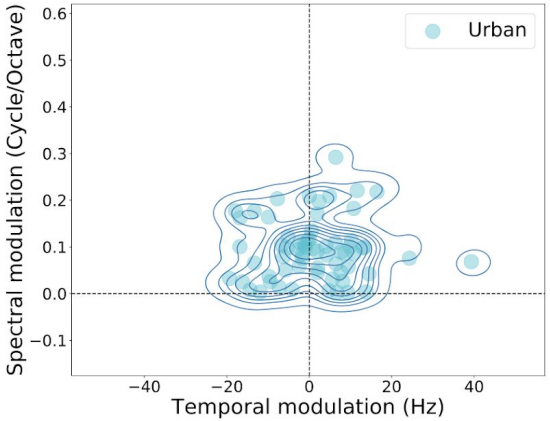
Voice Activity Detection	State-of-the-art on two datasets
Speaker Identification	Behind topline, above baseline
Bird Call type classification	Behind topline, above baseline
Audio scene classification	Close to topline

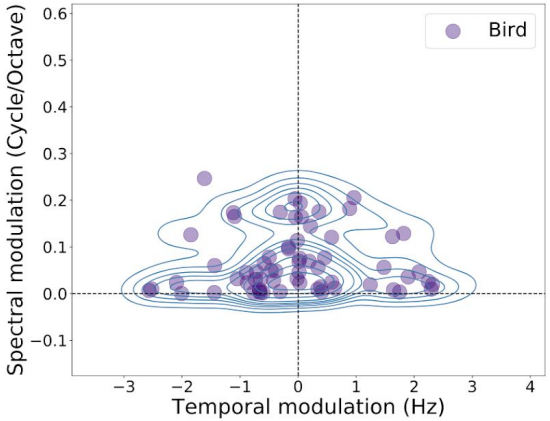
Qualitative analysis of filters



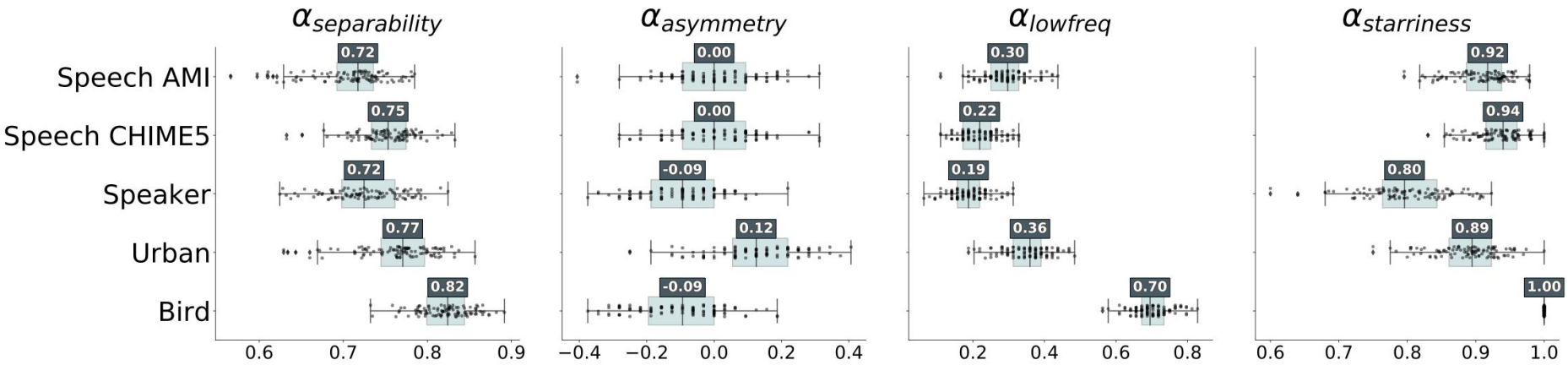






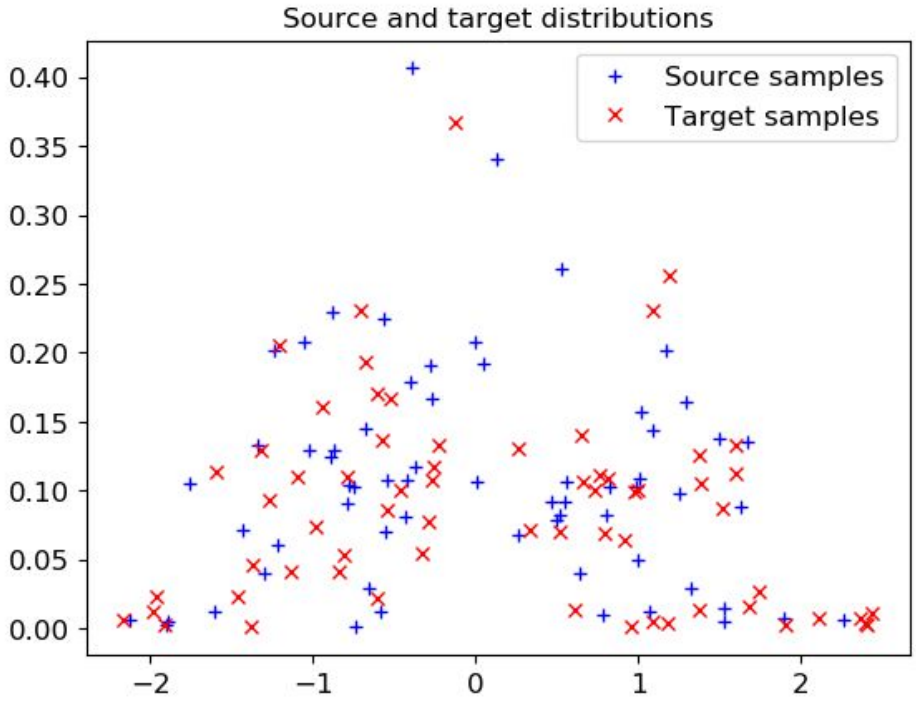


Quantitative analysis of filters



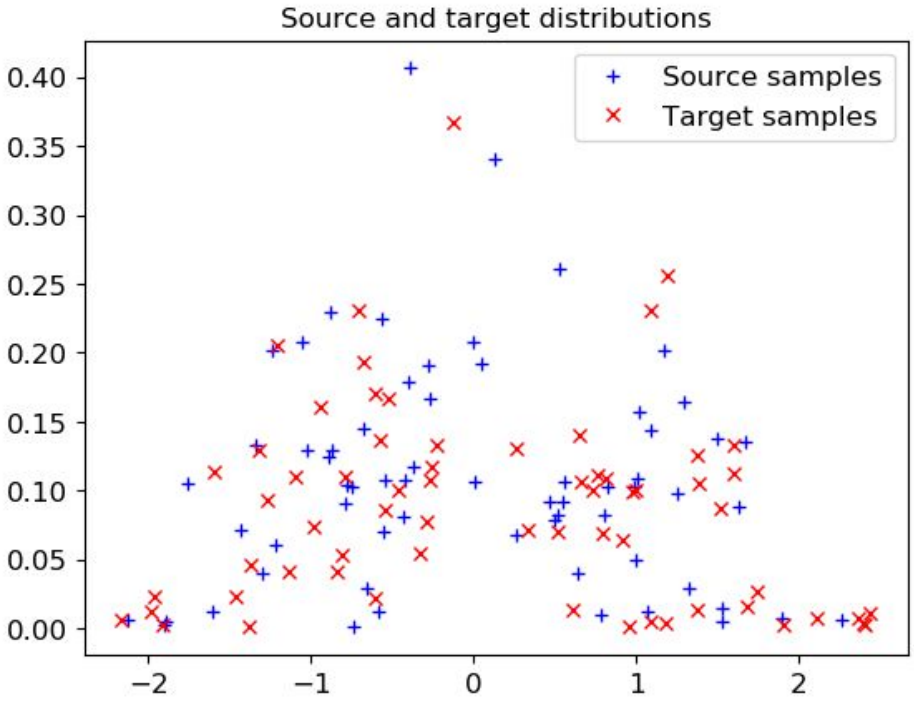
Quantitative Analysis of spectro-temporal

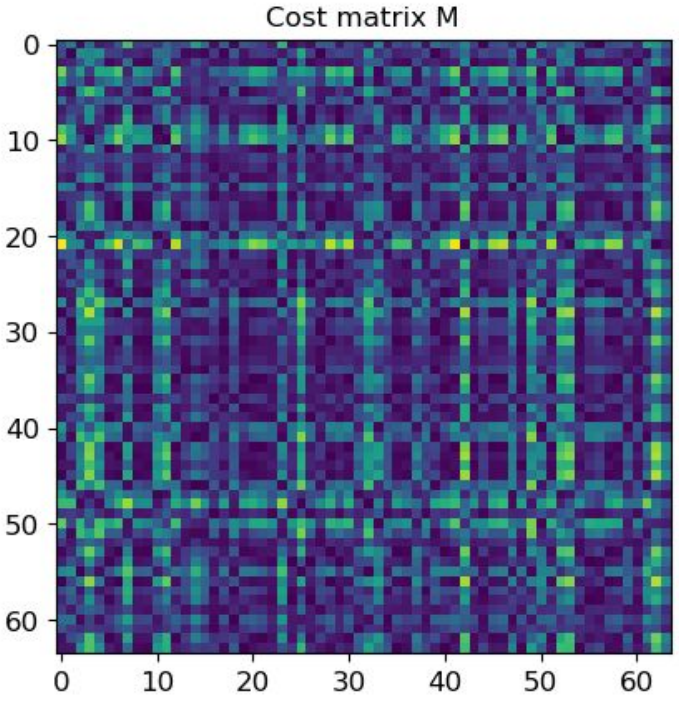
Distance between Learned STRFs with Optimal Transport



Quantitative Analysis of spectro-temporal

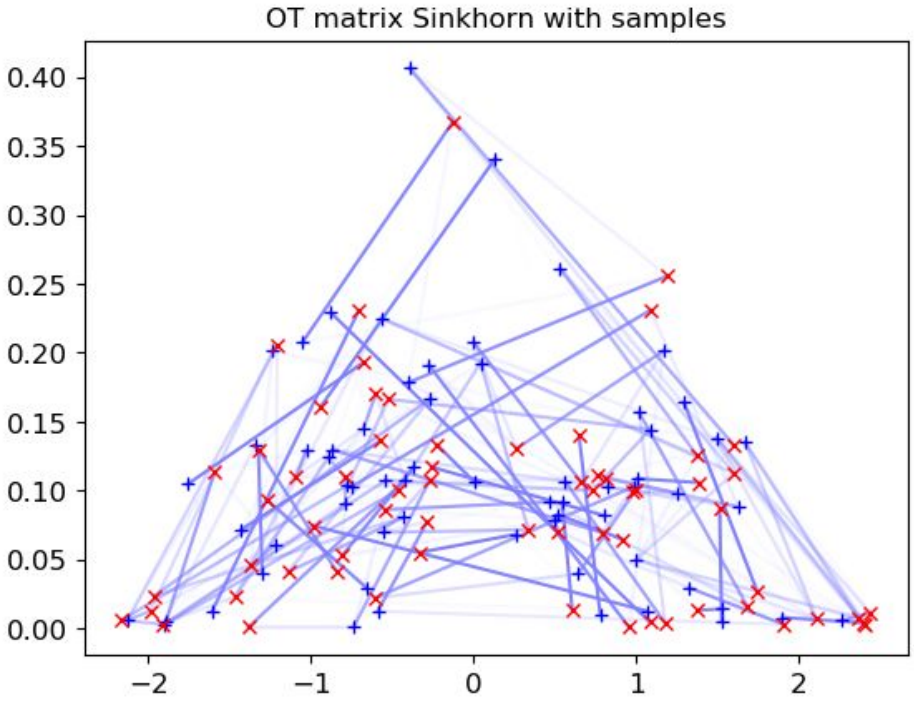
Distance between Learned STRFs with Optimal Transport

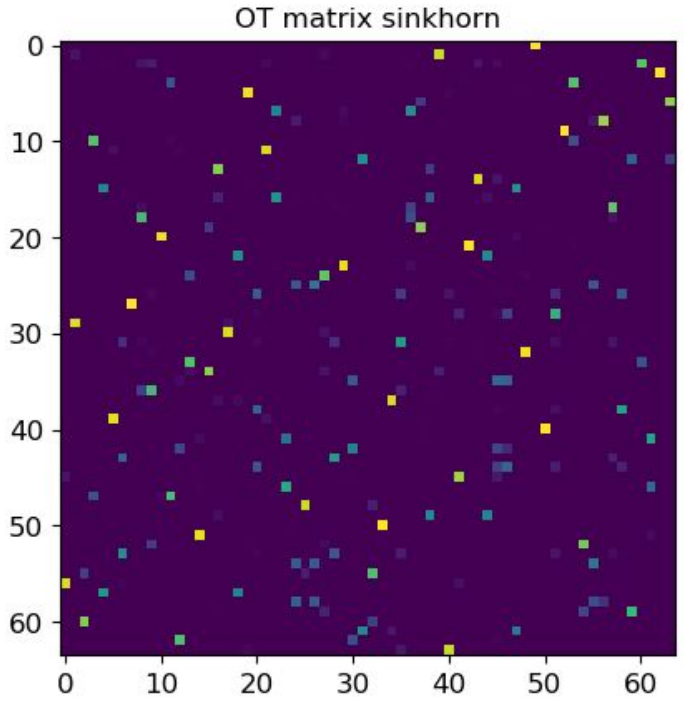




Quantitative Analysis of spectro-temporal

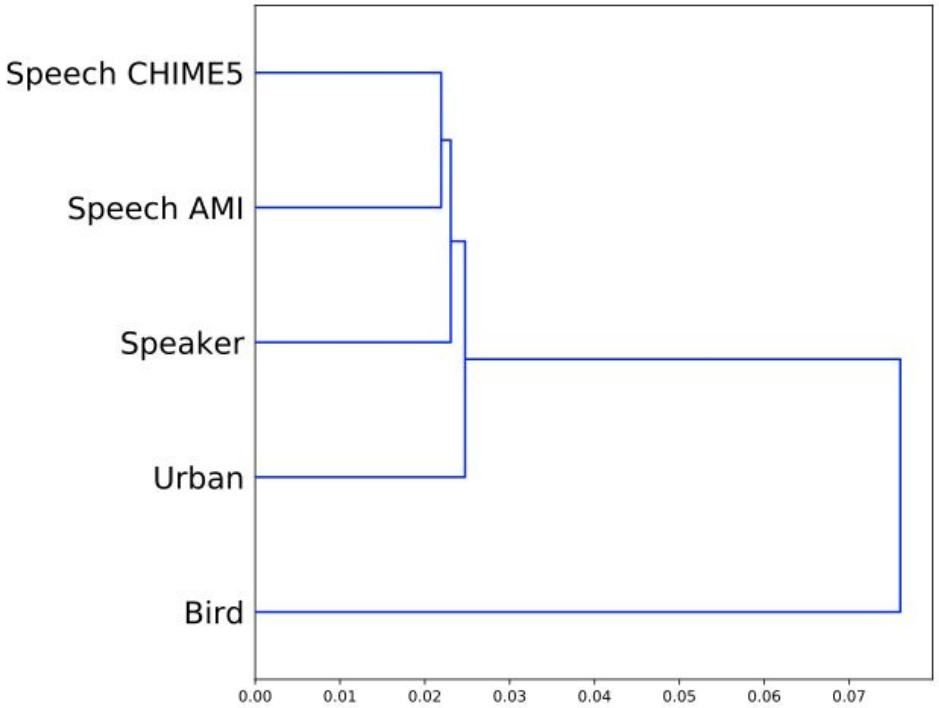
Distance between Learned STRFs with Optimal Transport





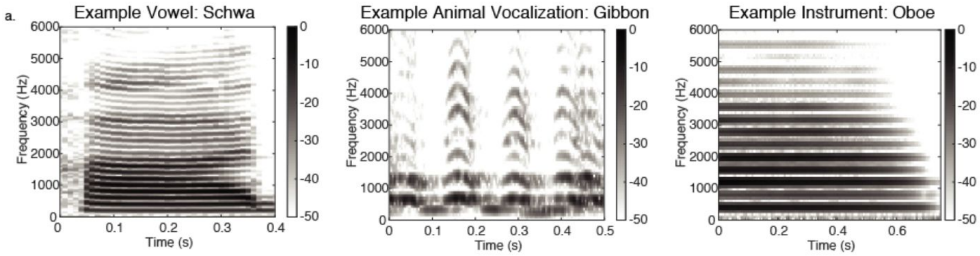
Quantitative analysis of filters

Distance between Learned STRFs with Optimal Transport



What is next?

1. Harmonicity



What is next?

1. Harmonicity

2. Tonotopic Organization for the Modulation

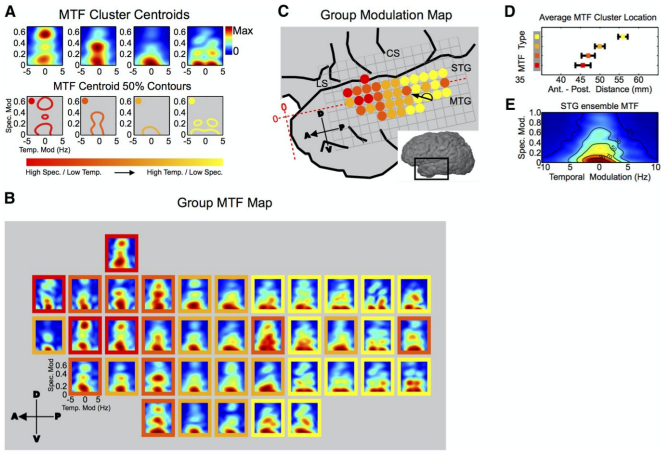


Figure 5. Organization of MTFs in STG. ***A***, K-means cluster centroids generated from all MTFs across all participants. Their respective 50% contours are shown below. The overall tuning within an individual MTF centroid is fairly well localized. The collection of MTF centroid types span modulation space from high spectral/low temporal regions (red, left) to high temporal/low spectral regions (yellow, right) as shown by the 50% contours. ***B***, Group MTF map. The map represents the average MTF across participants at each STG position. Only locations with ≥2 MTFs contributing to the average are included. Each MTF within the map is color-coded by its cluster membership. ***C***, MTF cluster identity map. MTF cluster identities from ***B*** are plotted. The cluster identity map shows a transition from high spectral/low temporal MTFs anteriorly (red) to high temporal/low spectral MTFs posteriorly (yellow) and a significant degree of local organization ($p < 1.0 \times 10^{-5}$, neighborhood similarity permutation test). ***D***, Average MTF cluster distance along the anterior–posterior extent of STG. Distances are measured from the anterior temporal pole illustrated in ***C*** (red horizontal axis). High spectral/low temporal MTFs are located anteriorly (red). High temporal/low spectral MTFs are located posteriorly (yellow, error bars represent SEM). ***E***, Ensemble MTF (population average) for STG. Contour lines represent percentage maximum. LS, lateral sulcus; STG, superior temporal gyrus; MTG, medial temporal gyrus; CS, central sulcus.