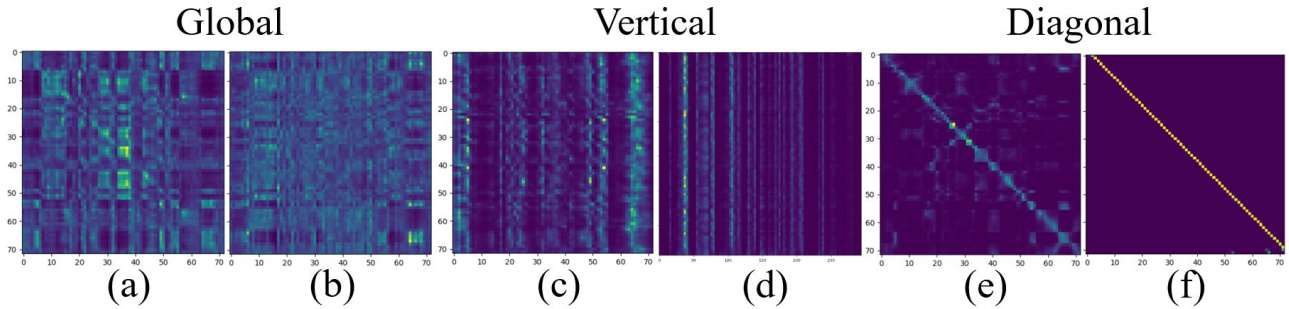
What do **self-attentions** learn

from **reconstruction** loss?

Attention categories

- Attention maps



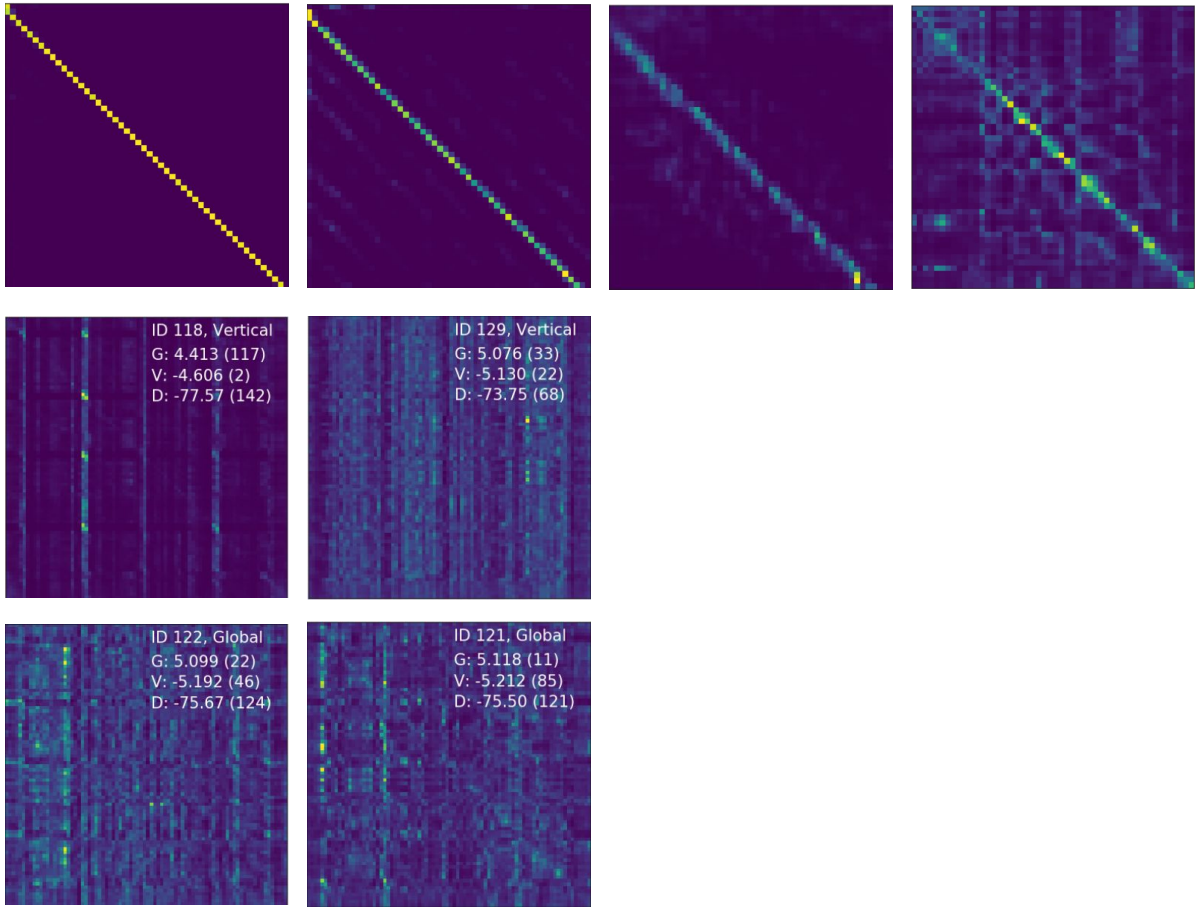
$$G(h) = \mathop{\mathbb{E}}_{u \sim U} \left[\frac{1}{T} \sum_{q=1}^{T} \mathbb{H}(A_u^h[q]) \right] \quad (1)$$

$$V(h) = \mathop{\mathbb{E}}_{u \sim U} \left[- \mathbb{H}(\frac{1}{T} \sum_{q=1}^{T} A_u^h[q]) \right] \quad (2)$$

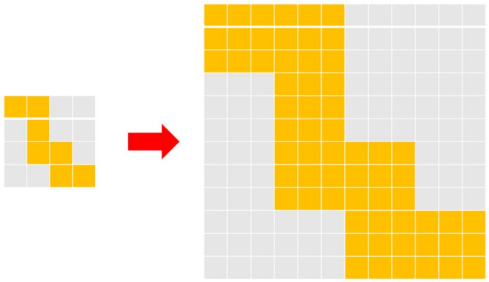
$$D(h) = \mathop{\mathbb{E}}_{u \sim U} \left[- \frac{1}{T^2} \sum_{q=1}^{T} \sum_{k=1}^{T} |q-k| \cdot A_u^h[q,k] \right] \quad (3)$$

Outline

- Diagonal
  - Focus
  - Block diagonal
- Vertical
- Global
  - Still working on



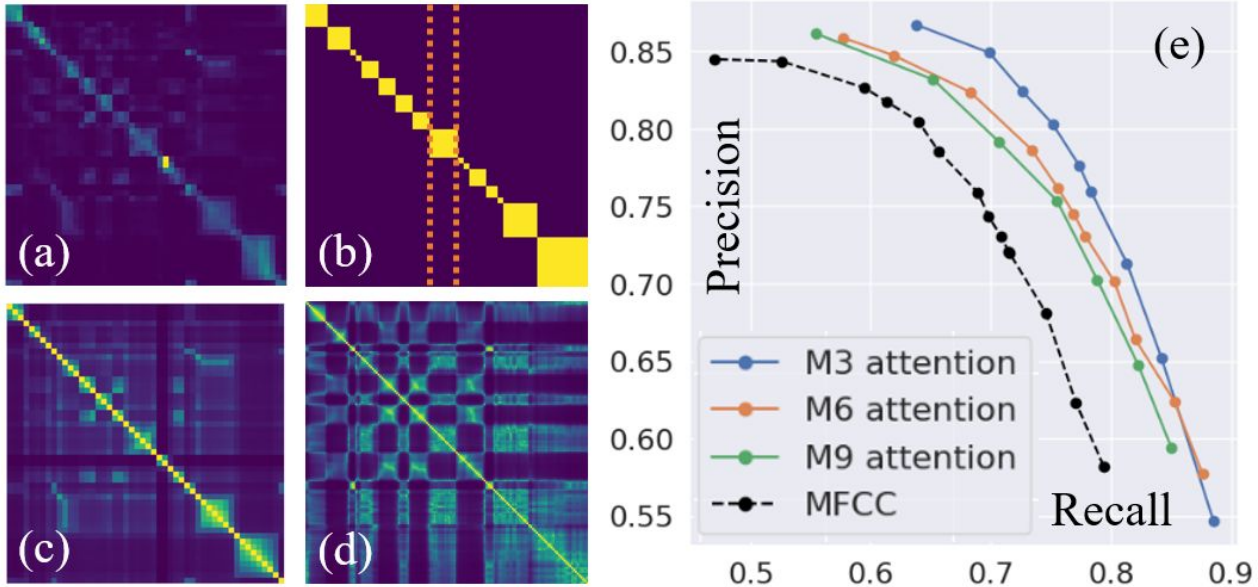
Phoneme segmentation



True boundaries

Attention map

Similarity matrix on (a)

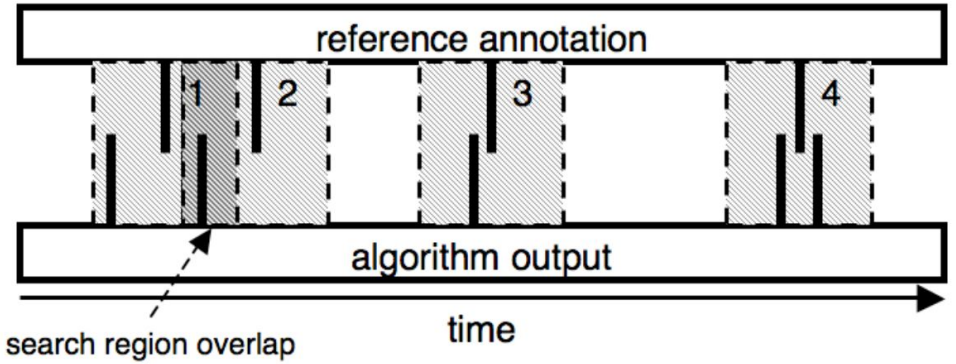


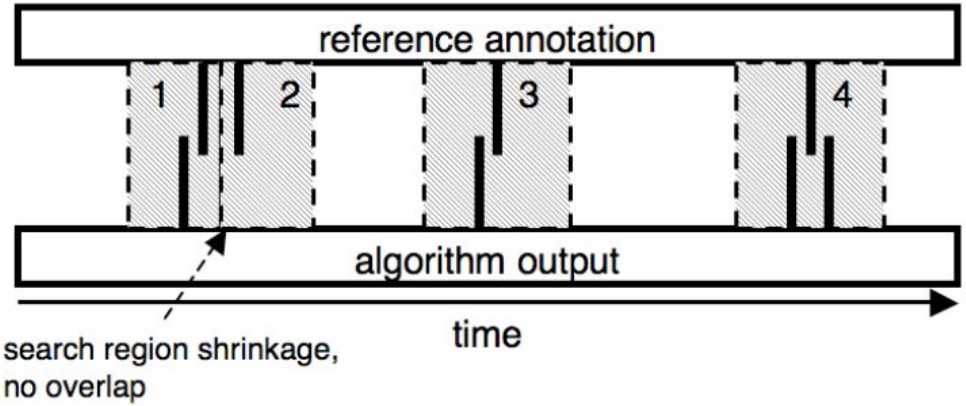
Similarity matrix on MFCC

$$K[i,j] = exp(-\frac{|| v_i - v_j ||}{\alpha})$$

Evaluation: Precision-recall

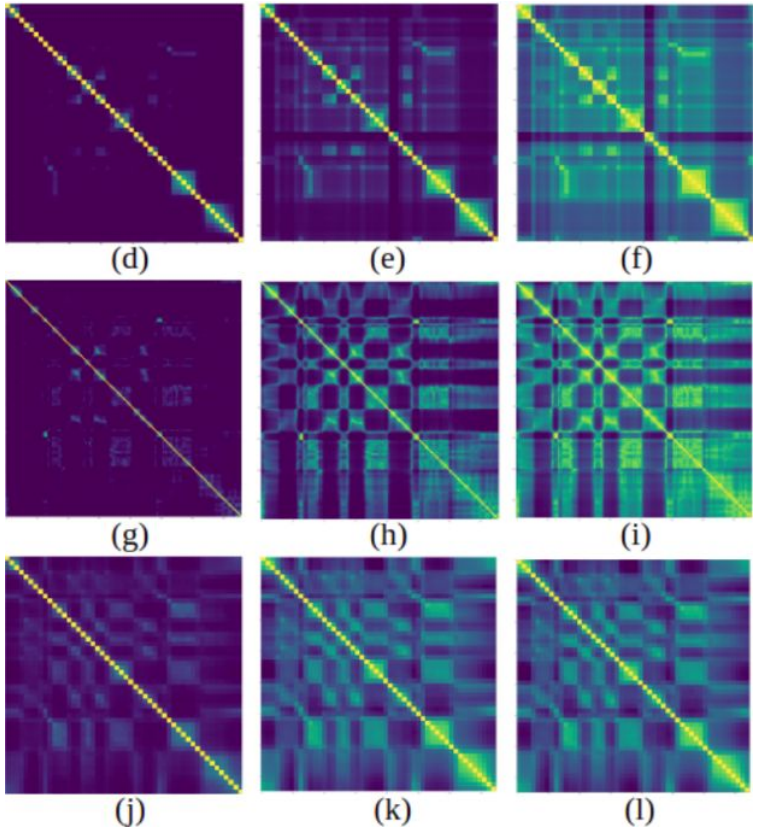
- How to count a hit?





Block diagonal learned to neglect

- Further frames → more different? NO



Attention map

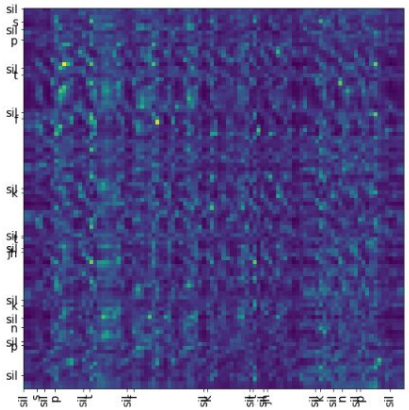
MFCC

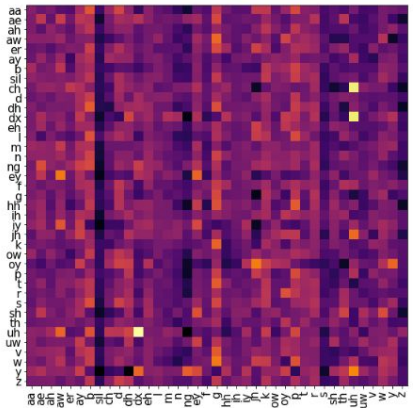
Representation

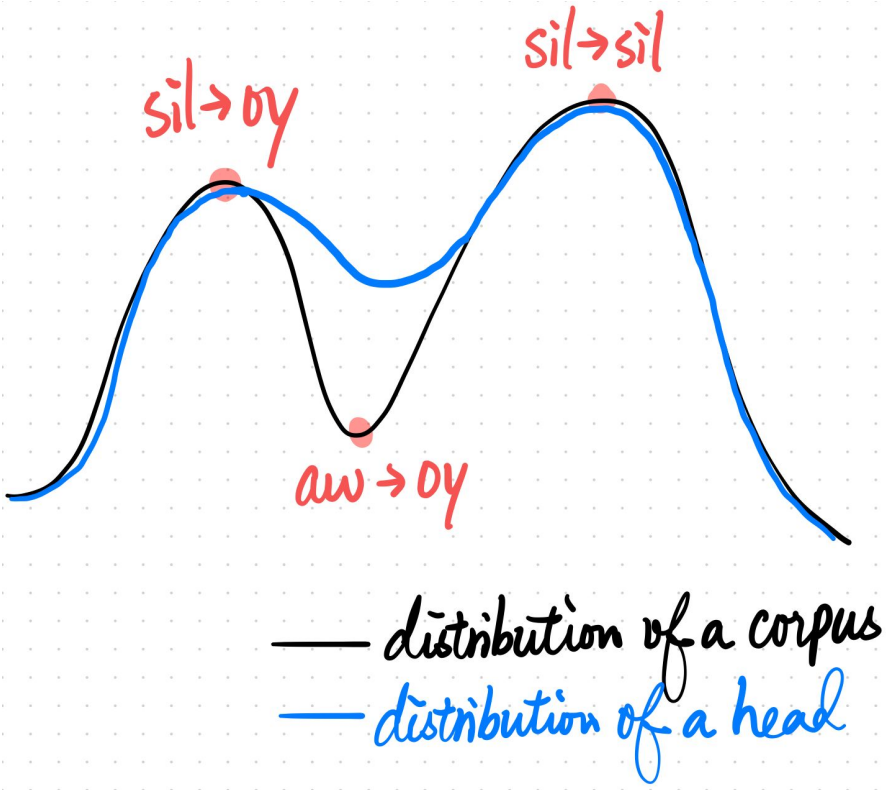
Phoneme relation map

- Different input utterance → different attention map
- How to summarize the operation of a head (if exists)
  - Propose to align to phonemes

$$P_h^{'}[m,n] = \mathop{\mathbb{E}}_{u \sim U}\left[\frac{1}{T}\sum_{q=1}^{T}\sum_{k=1}^{T}\mathbb{I}_{y_q=Y_m}\cdot\mathbb{I}_{y_k=Y_n}\cdot A_u^h[q,k]\right]$$

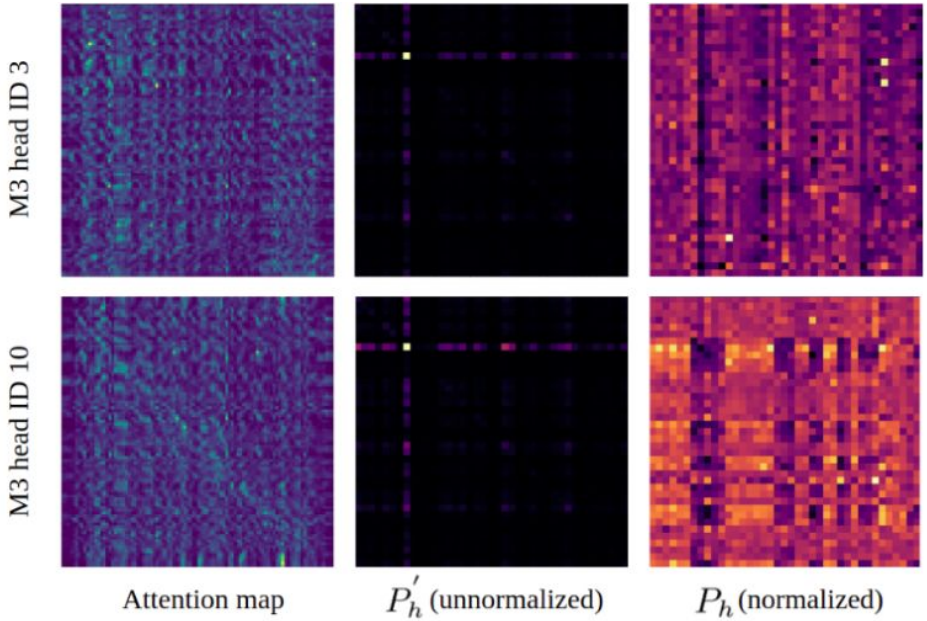


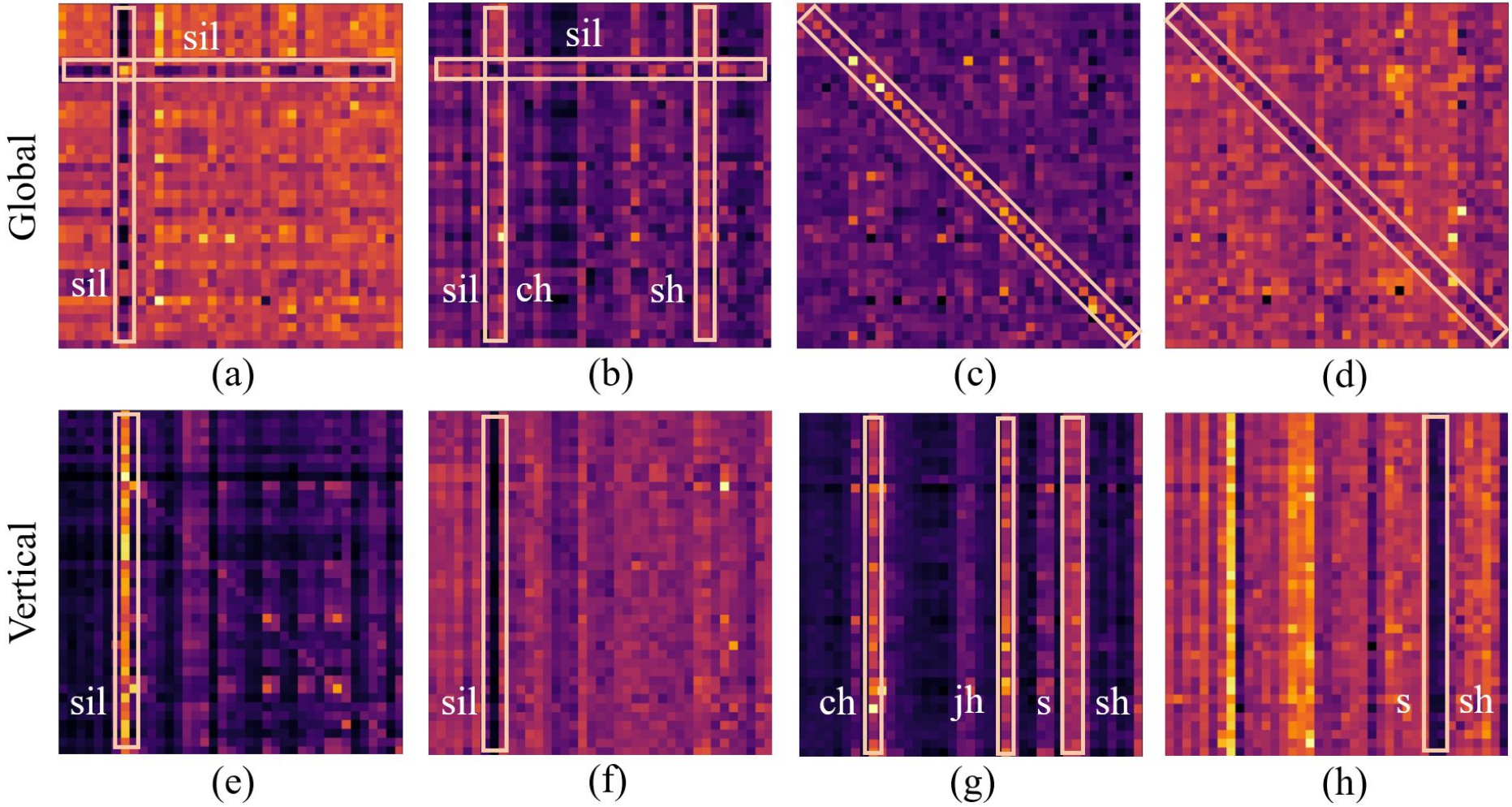


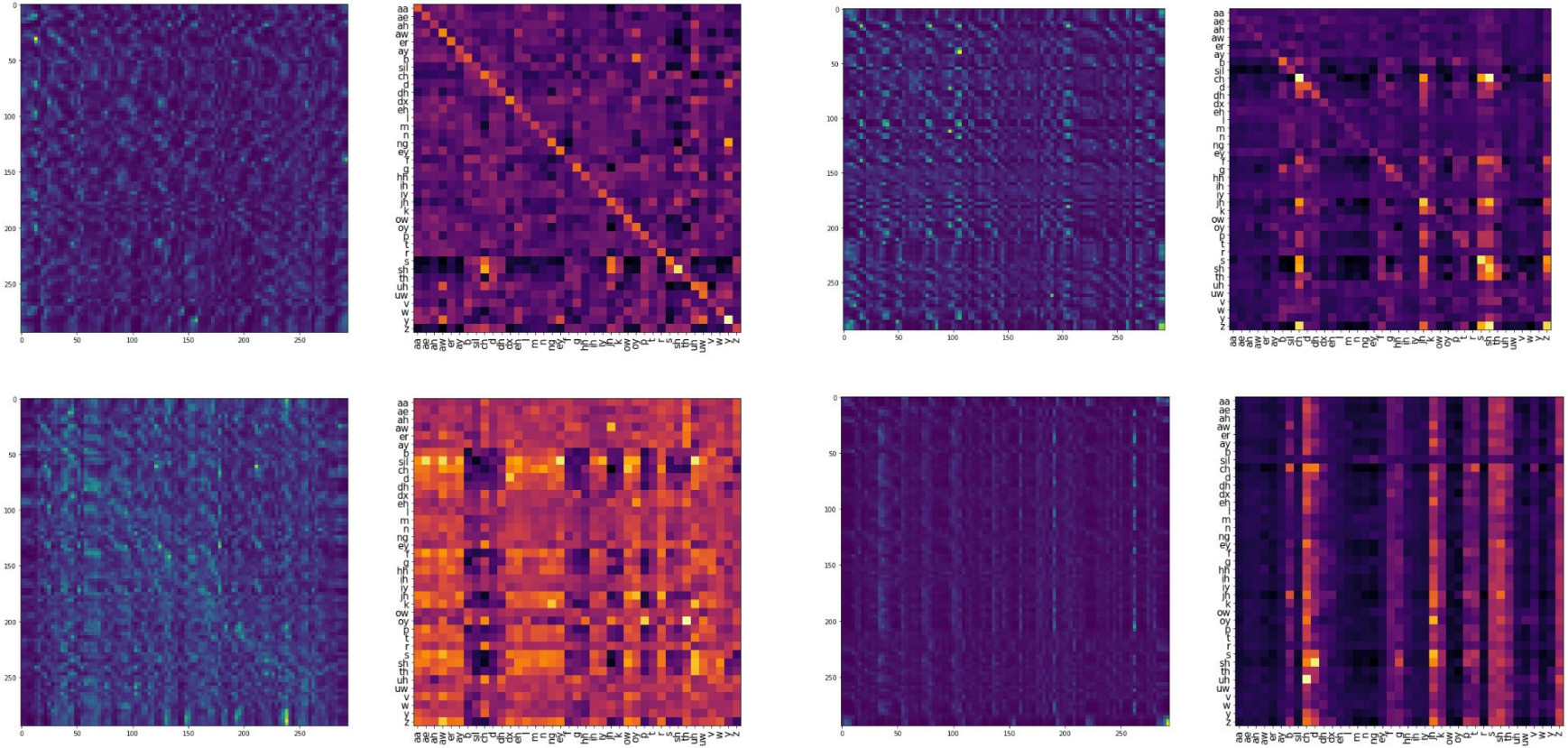


$$P_h[m,n] = \frac{P_h^{'}[m,n] - P_U[m,n]}{P_U[m,n]}$$

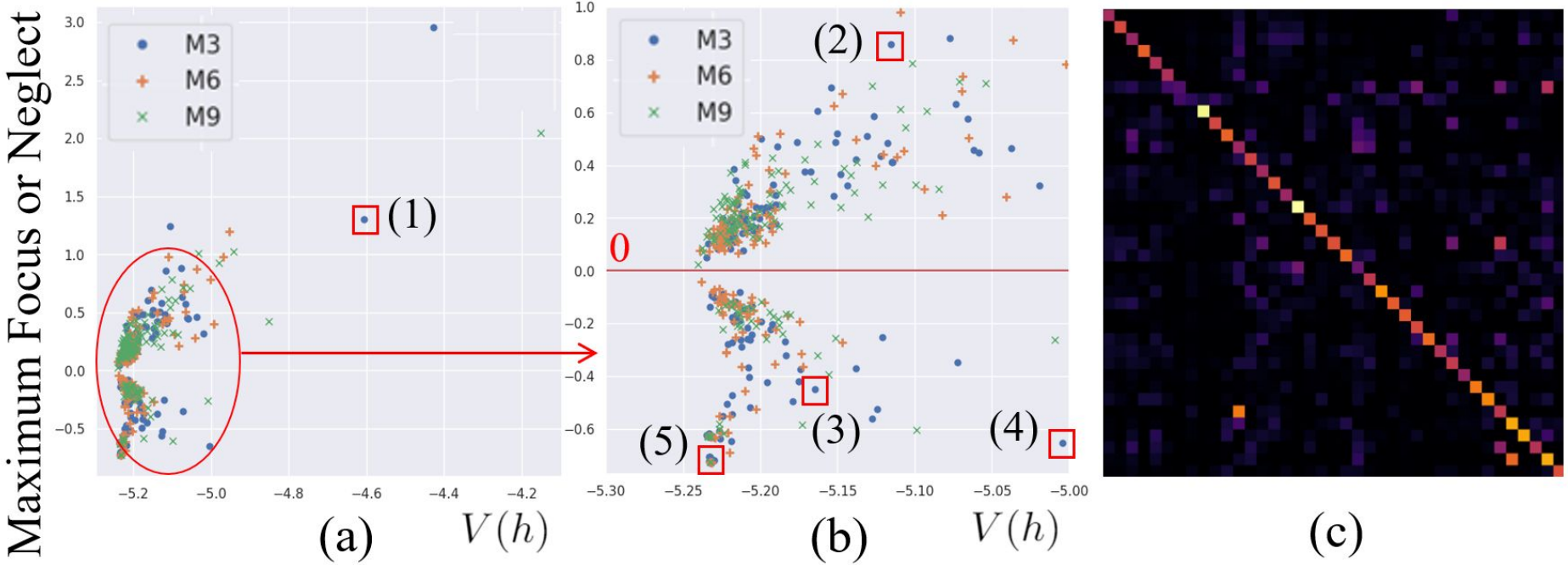
- The positive represent preference




Global
Vertical
sil
sil
sil
sil
ch
sh
(a)
(b)
(c)
(d)
sil
sil
ch
jh
s
sh
s
sh
(e)
(f)
(g)
(h)

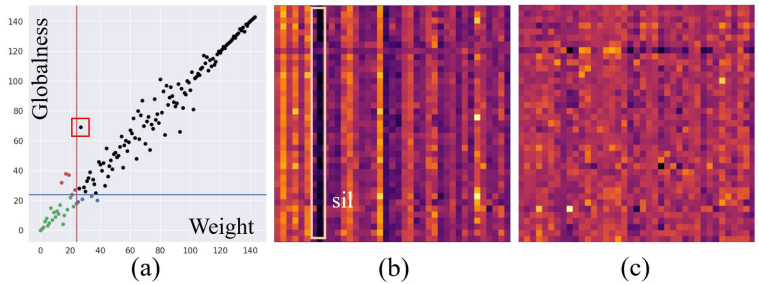


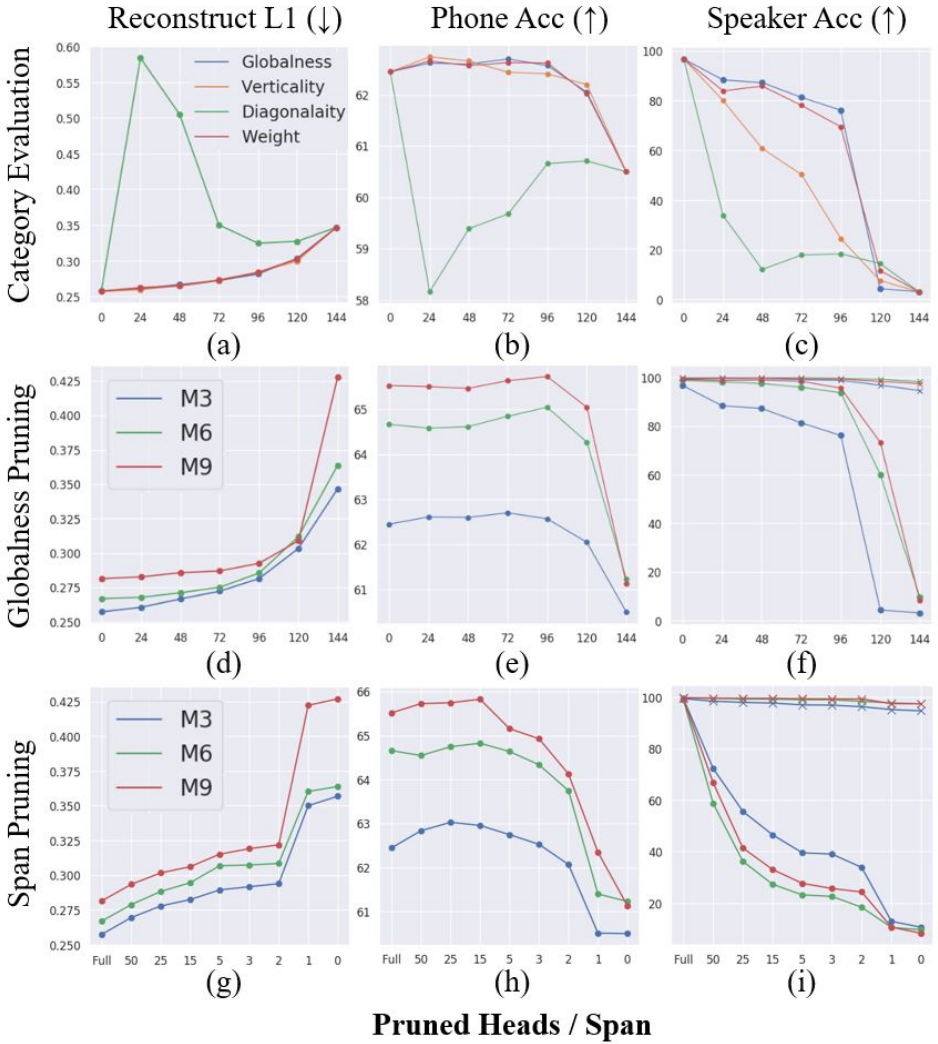
$$C_h[n] = \frac{1}{|Y|} \sum_{m=1}^{|Y|} P_h[m, n]$$



Importance ranking

- Diagonal > Vertical > Global
- Compare with **weight**





Take-away

- Reconstruction:
  - Phonetic information → Diagonal attentions
    - Aware of pretraining mask length
    - Phoneme interval
  - Speaker identity → Vertical attentions
    - Focus
    - Neglect