Representational analysis in Self-Supervised Speech Models

Decoding Speech Representations: Cognitive Connections and Conceptual Challenges

81a9c	32	3D	4B	70	B7	5B	EF	53	E1	38	EA	40	2A	5E	D2	79	DF	D2	0E	21	CF	88	BA
81acb	3B	7A	CF	3E	DB	7D	31	8D	99	88	04	E1	8D	1C	2D	6D	3B	22	37	70	4A	8D	BF
81afa	8F	CD	2E	1D	8A	9F	BC	3F	50	EF	47	E5	4E	84	2D	C6	09	79	52	4A	77	22	07
81b29	49	B5	34	B2	2B	53	E0	97	06	E4	EE	22	3D	FD	B1	E9	F8	72	B0	62	EE	EE	BC
81b58	13	8E	6F	5F	73	21	0D	7F	BA	D8	17	14	6D	25	5E	7A	91	72	6C	59	D9	BA	69
81b87	E3	23	3A	AC	EA	A6	A0	55	D2	7C	4D	0A	3C	CB	71	63	58	E2	26	49	3F	94	63
81bb6	27	CA	9A	74	21	64	A7	68	09	9D	C9	FA	1F	BE	38	5D	77	05	90	63	CE	F3	F5
81be5	54	7F	48	38	E6	30	5A	D7	39	AD	6F	52	79	5D	04	D3	BE	3C	27	16	F5	A5	52
81c14	27	B0	05	B2	3E	F8	F4	A8	08	C0	CB	82	31	D1	E4	EE	BF	A7	65	C8	E3	63	0C
81c43	A8	CB	74	4D	78	31	85	C9	C1	8D	34	7A	93	A2	AF	4F	2B	D1	3F	87	1A	52	C6
81c72	B0	F8	47	1D	D7	A5	E8	B1	B9	B0	ED	BE	13	81	96	A8	FA	65	9B	AE	75	CF	B4
81ca1	20	C9	8B	D3	9B	C6	6B	5E	63	C8	F7	65	22	8F	42	5A	44	84	90	21	49	DC	1E
81cd0	1A	9B	5D	ED	A3	69	A9	65	B7	C2	54	15	A2	24	09	DE	67	D7	DB	91	38	BF	9E
81cff	CB	E8	43	5E	2D	59	D6	DA	76	48	2A	52	47	1D	80	27	0D	7E	B0	3F	D3	DA	D7
81d2e	09	FD	FA	6C	4D	78	44	27	85	E9	00	C7	E4	71	C7	F8	2F	16	4C	DD	4B	22	BA
81d5d	CB	4C	A8	3E	52	BE	55	CE	DE	BB	E3	D4	F0	B0	43	6E	27	F4	0B	87	D5	32	24
81d8c	51	9F	B9	02	7D	B1	D3	45	83	17	95	BD	70	8F	CB	91	D3	9A	3D	57	A0	F2	A6
81dbb	63	8E	D5	1F	1C	99	1B	01	5D	96	81	2C	98	63	CC	0B	09	EA	46	6E	AE	46	7A
81dea	AF	8C	35	19	4E	A8	25	8C	F6	0A	53	E0	6D	3D	49	B4	37	5F	67	A8	02	B6	DC
81e19	99	80	FD	A5	E8	DE	8A	E4	24	14	7E	D3	D1	25	2C	A4	13	C1	29	D3	09	3E	D3
81e48	56	CC	EA	AA	57	9E	0D	8A	67	11	AD	71	04	05	7A	8F	4F	FB	B1	DF	66	E3	9C

(a) hex dump of picture of a lion

(b) same lion in human-readable format

Figure 1: The hex dump represented at the left has more information contents than the image at the right. Only one of them can be processed by the human brain in time to save their lives. Computational convenience matters. Not just entropy.

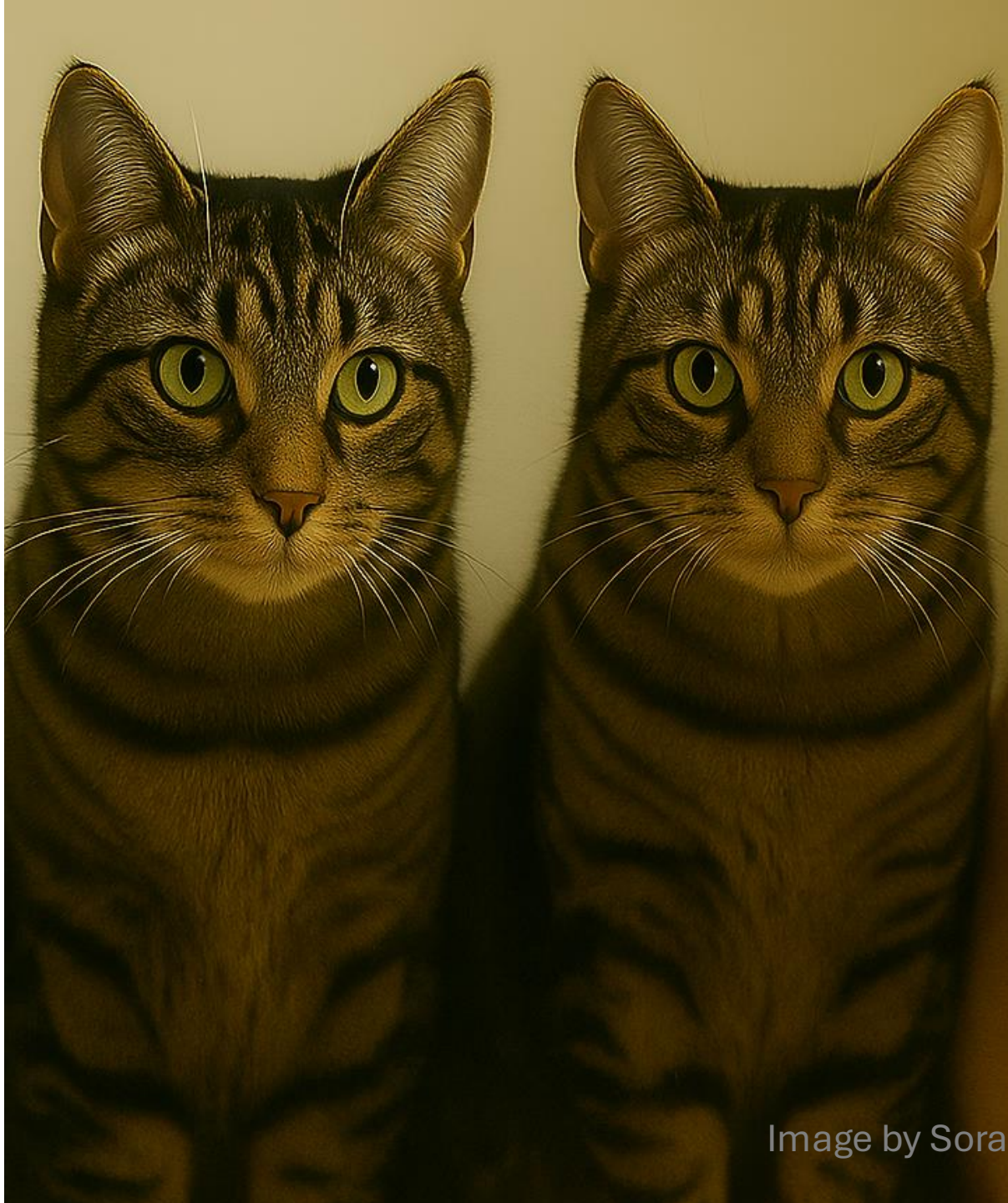
Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

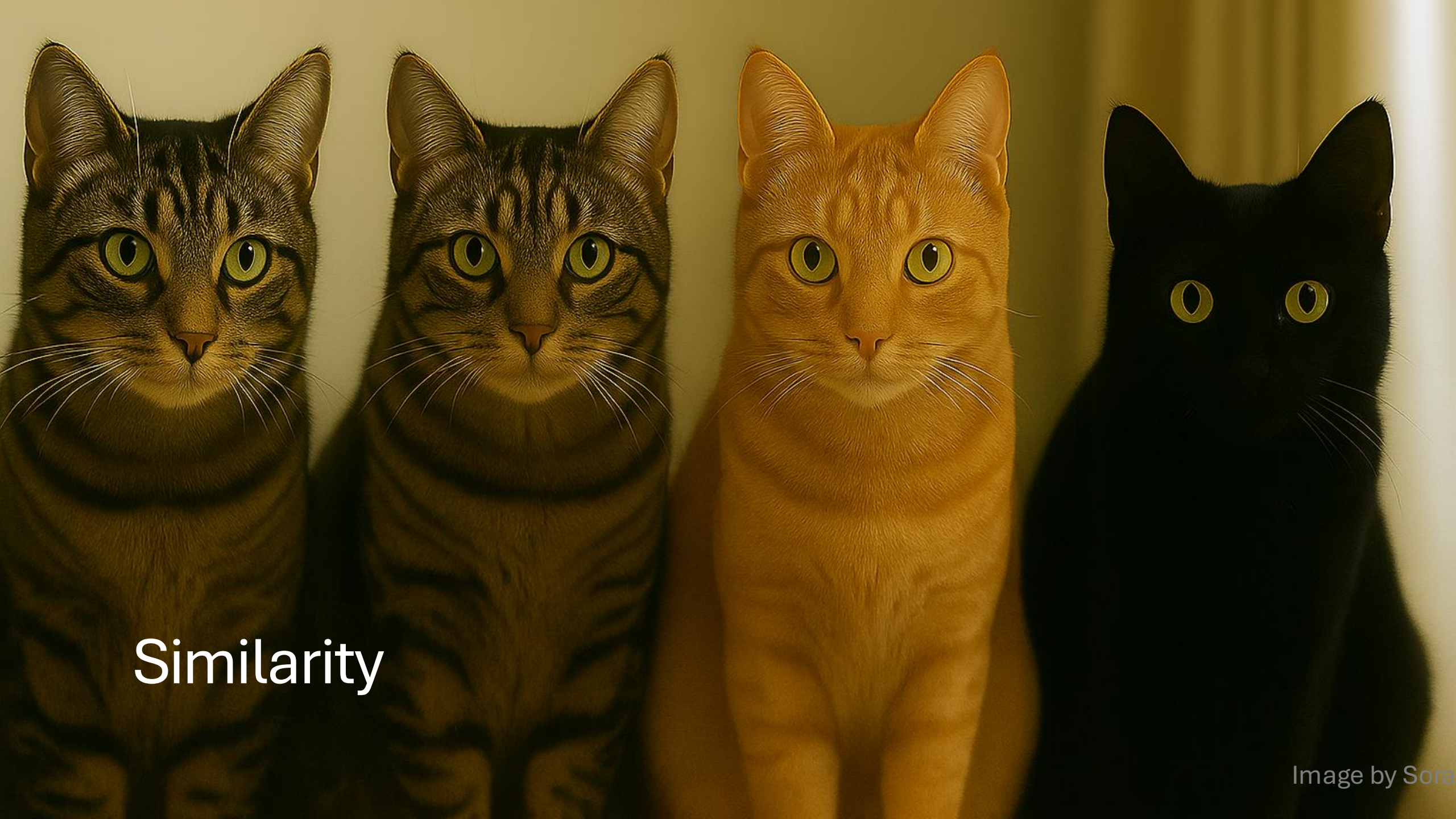
Representation

/ k æ t s /
[k^h æ t s]

cat (root morpheme)
-s (suffix morpheme)

cats (plural noun)





Similarity

Image by Sora

Similarity & embeddings

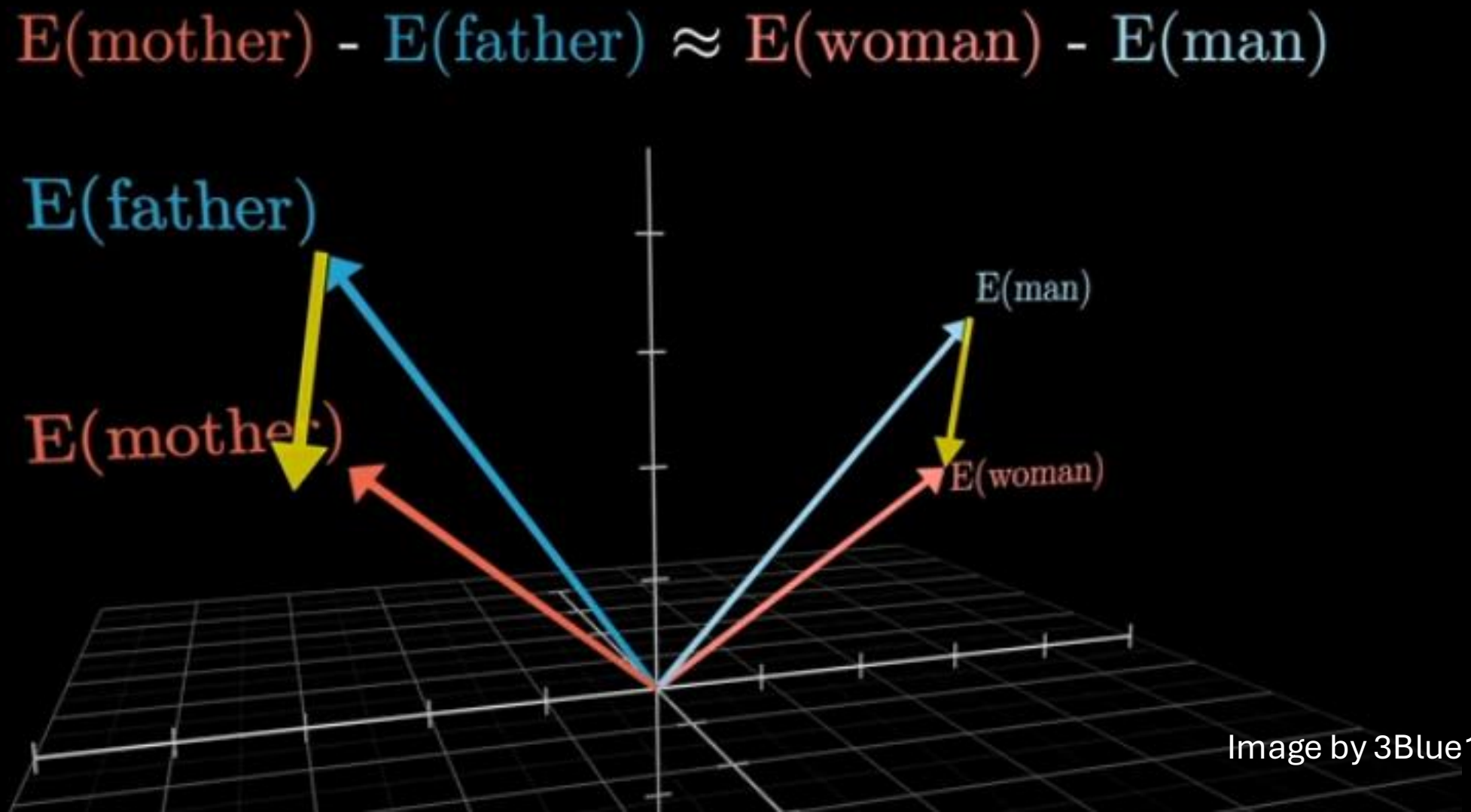
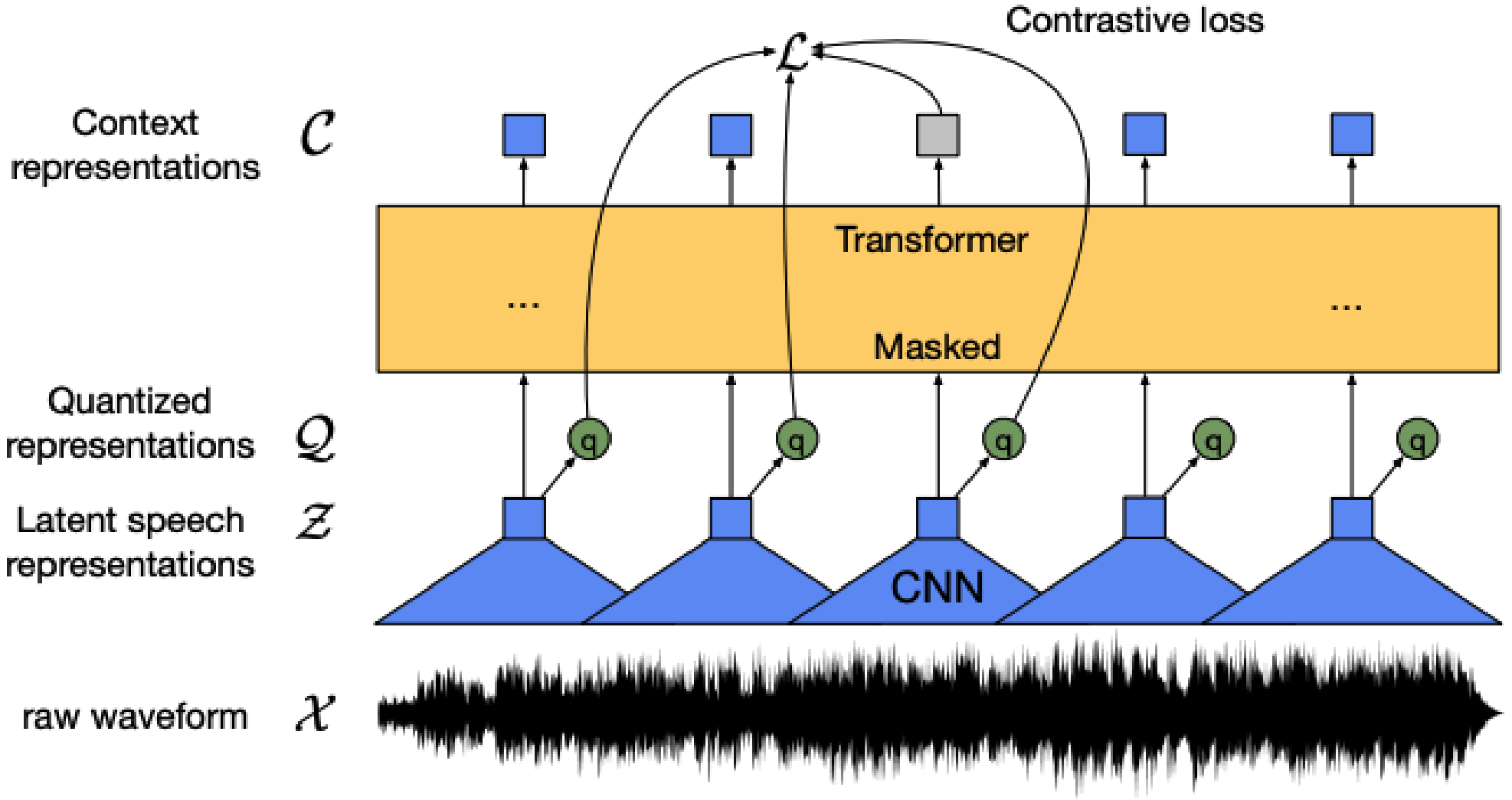


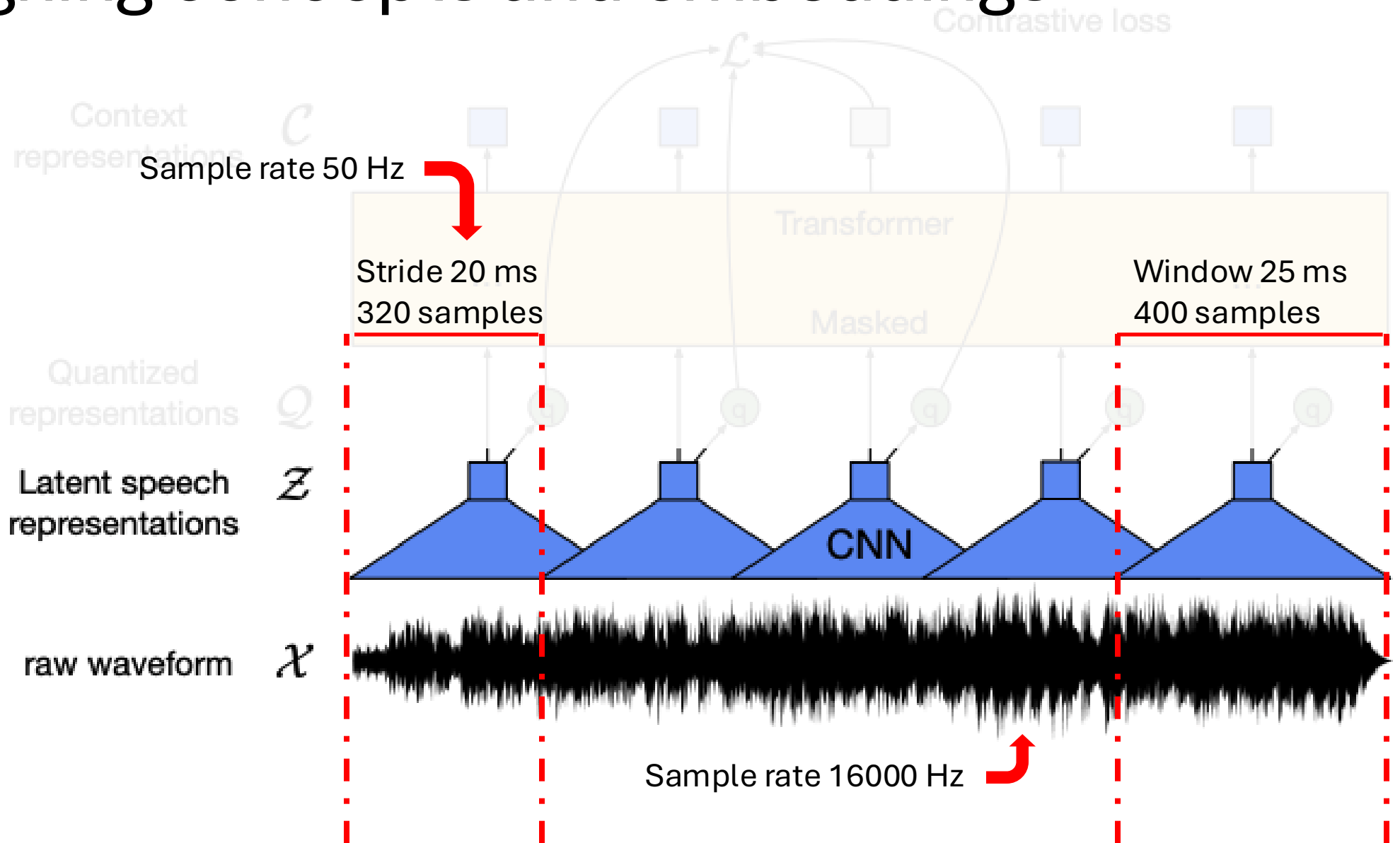
Image by 3Blue1Brown

Wav2Vec 2.0



Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Aligning concepts and embeddings



Aligning concepts and embeddings

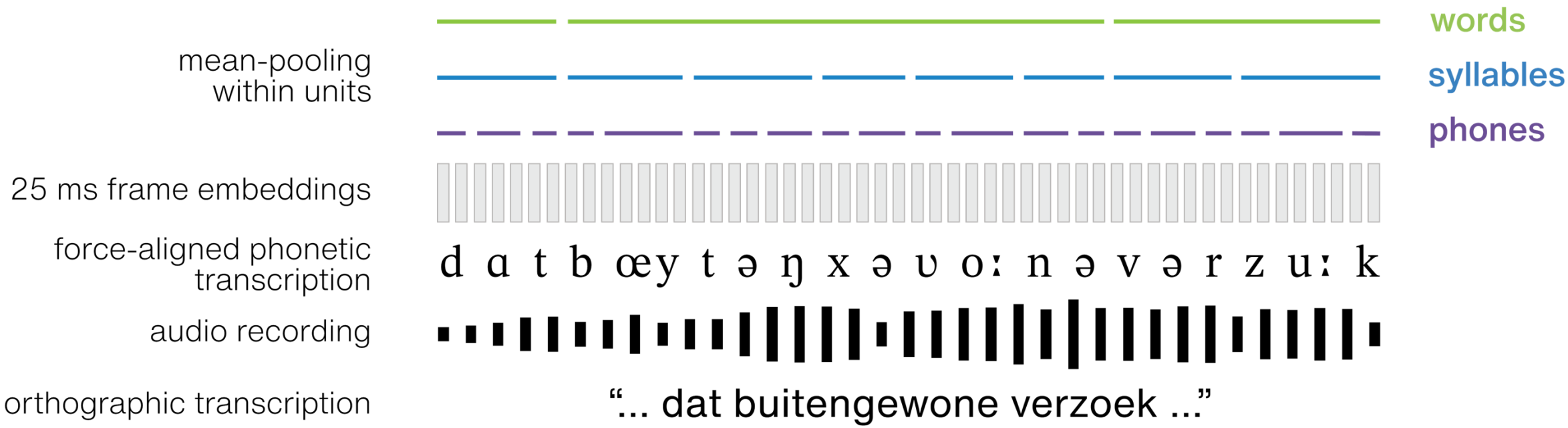
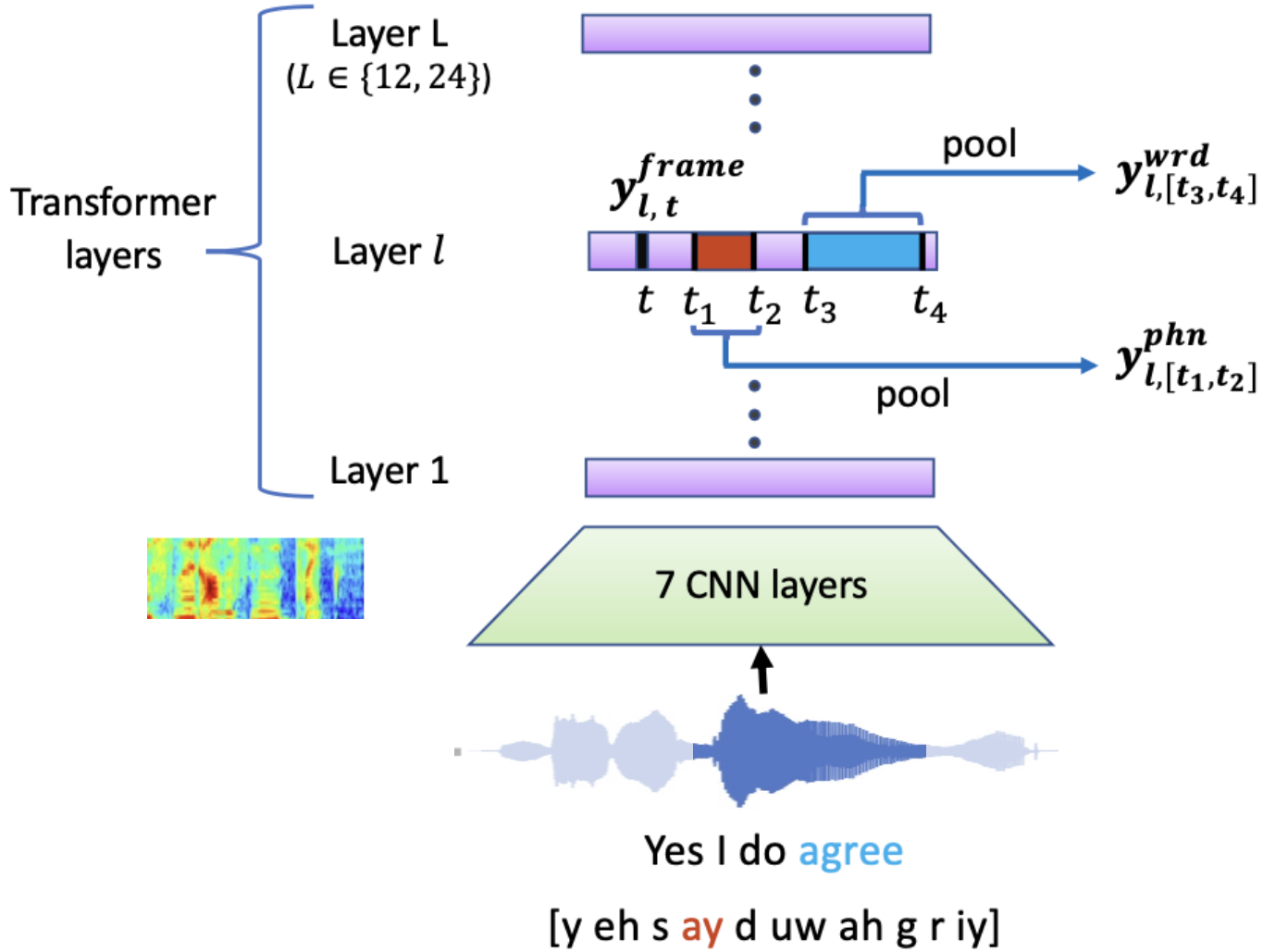


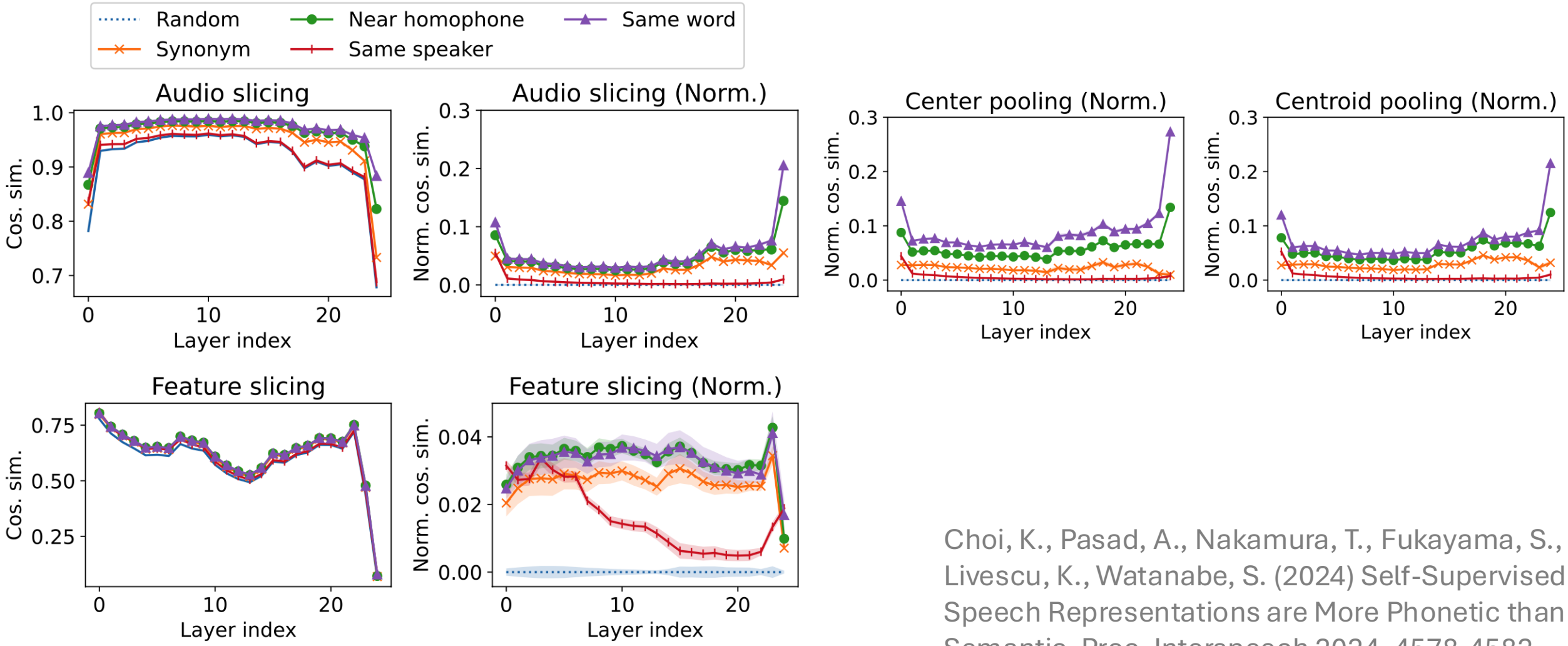
Image by Marianne de Heer Kloots

Aligning concepts and embeddings



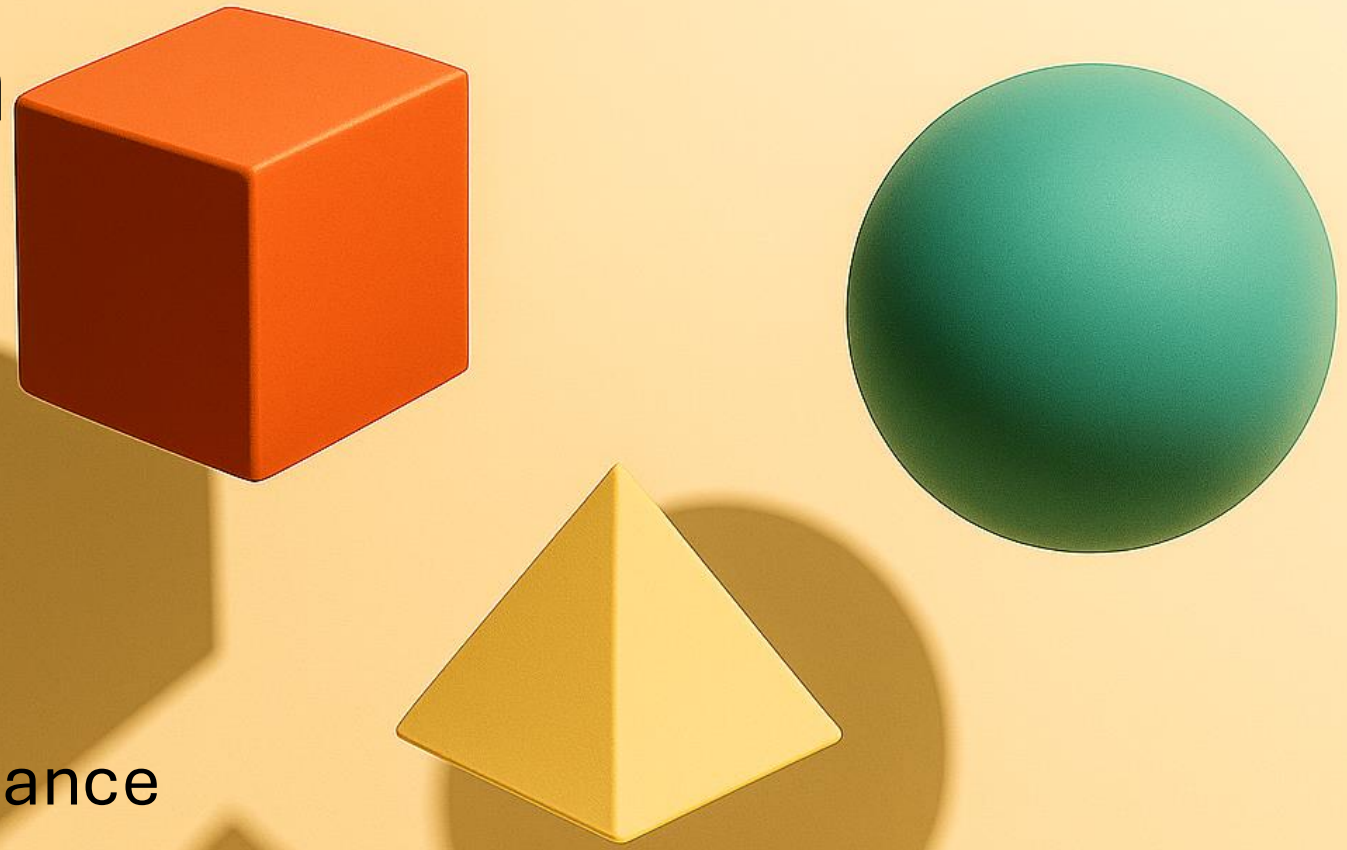
Pasad, A., Chou, J. C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 914-921). IEEE.

Aligning concepts and embeddings



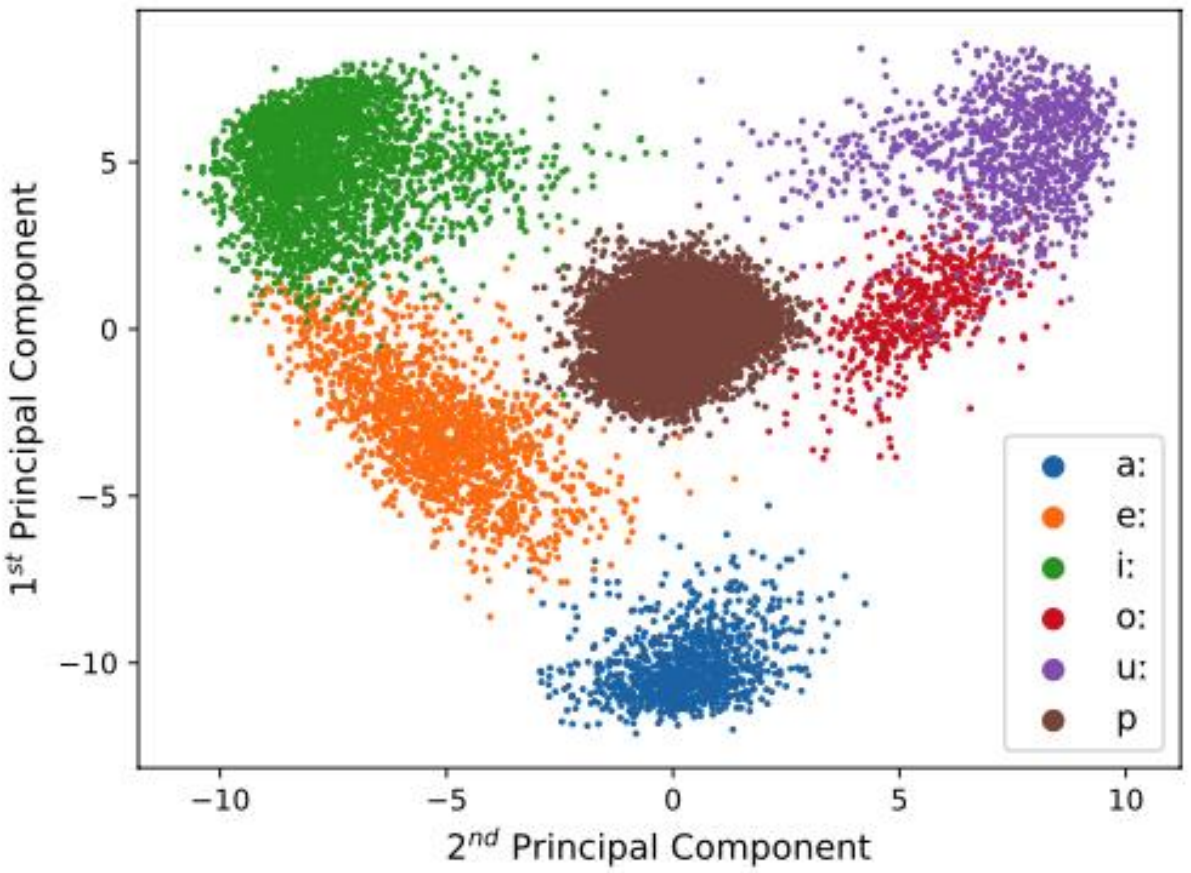
Choi, K., Pasad, A., Nakamura, T., Fukayama, S., Livescu, K., Watanabe, S. (2024) Self-Supervised Speech Representations are More Phonetic than Semantic. Proc. Interspeech 2024, 4578-4582

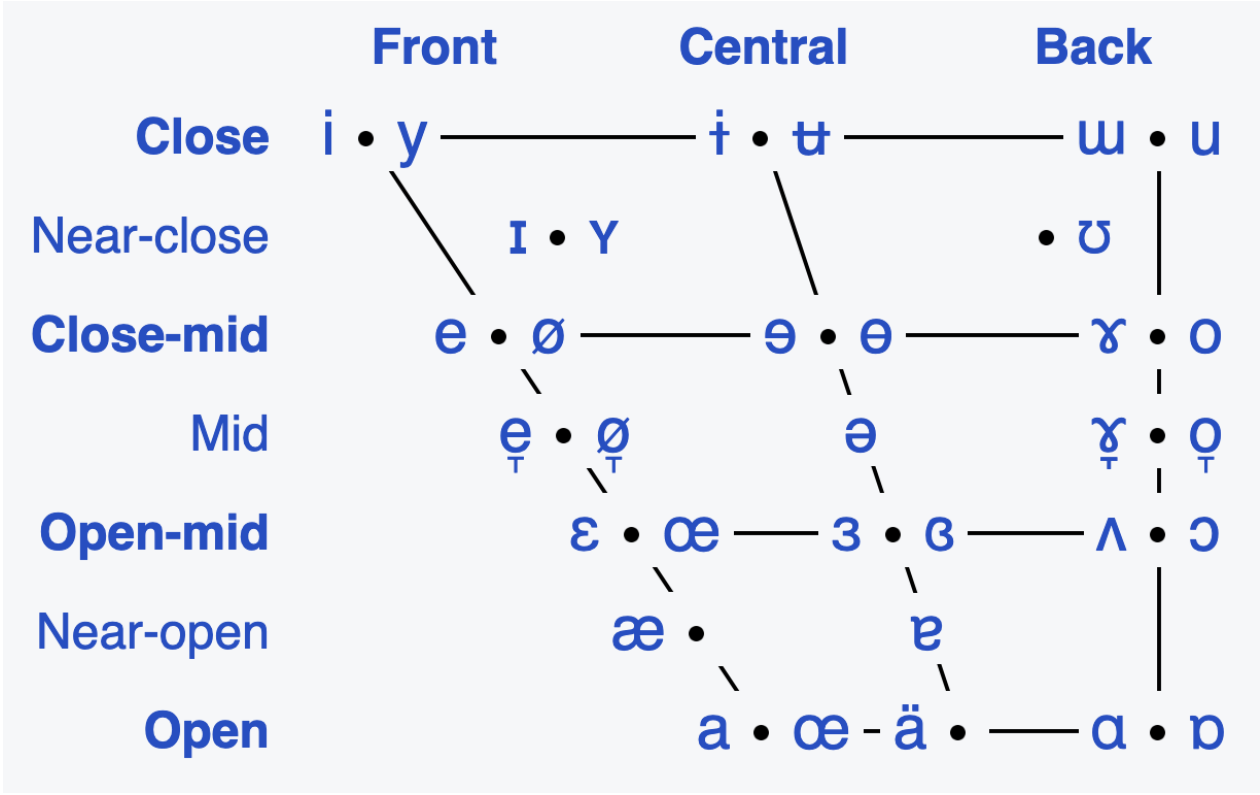
Dimension reduction



- PCA — based on variance
- LDA — based on class discriminability (supervised)
- MDS — based on pairwise distances
- t-SNE — based on probability distributions of pairwise distances
- UMAP — based on local and global structure of pairwise distances

Dimension reduction

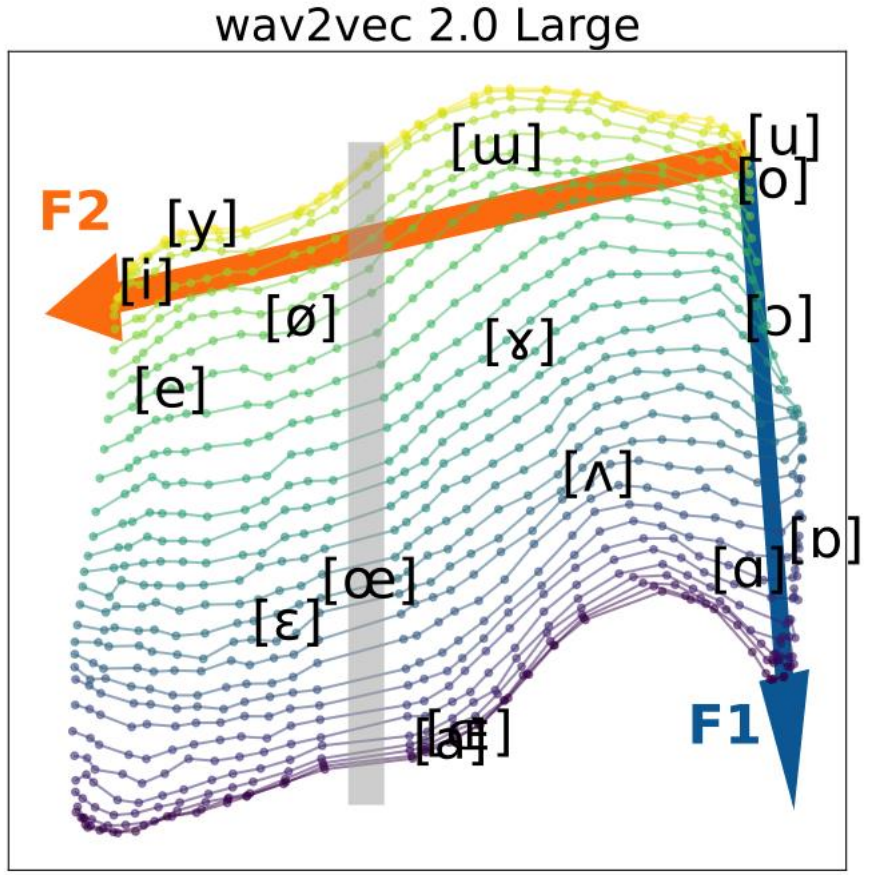


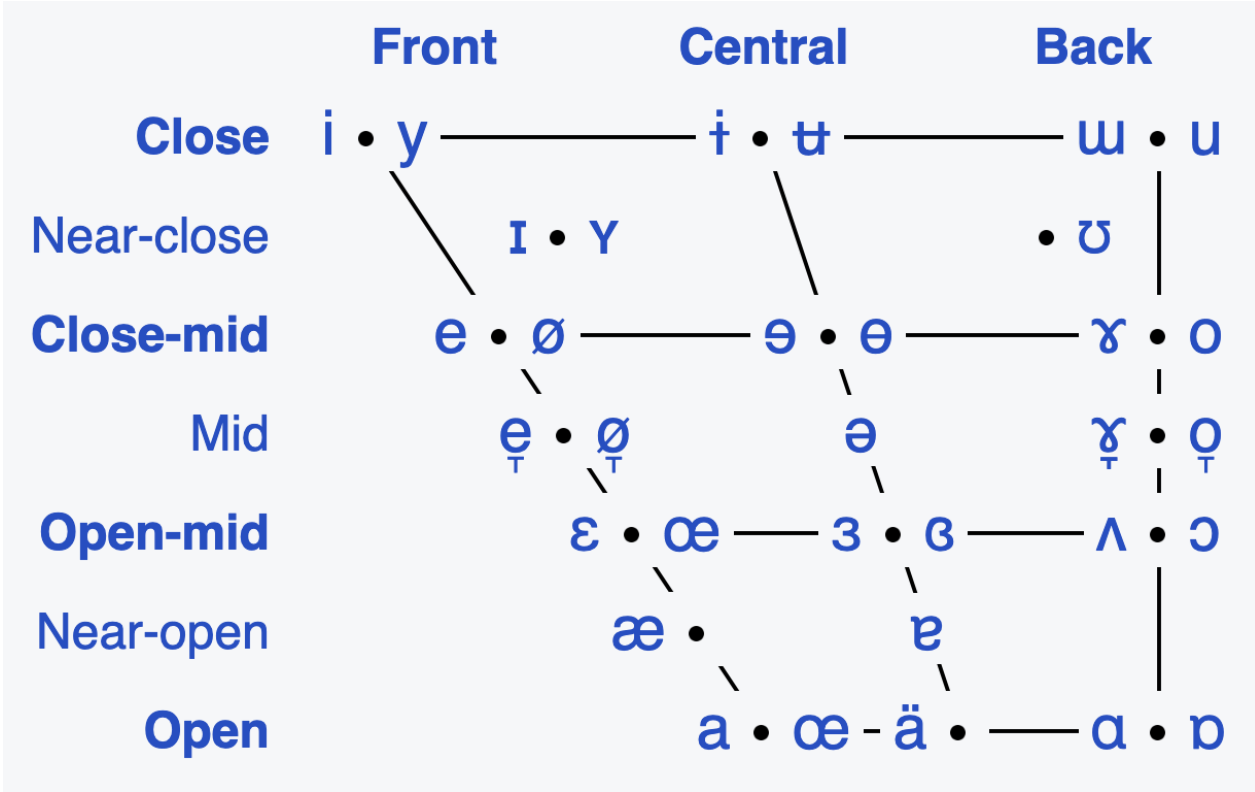


Wikipedia

Dieck, T., Pérez-Toro, P. A., Arias, T., Nöth, E., & Klumpp, P. (2022). Wav2vec behind the Scenes: How end2end Models learn Phonetics. In *Interspeech* (pp. 5130-5134).

Dimension reduction





Choi, K., & Yeo, E. J. (2022). Opening the black box of wav2vec feature encoder. *arXiv preprint arXiv:2210.15386*.

Wikipedia

Pretrained on ambient sounds

Poli, M., Schatz, T., Dupoux, E. & Lavechin, M. (2025). Modeling the initial state of early phonetic learning in infants. Language Development Research, 5(1), 1-34.

Baseline and controls

- Input pertubation, context selection
- Model trained on (non-)speech, randomly initialized, baseline model, baseline features
- Probe ABX, diagnostic classifier, RSA
- Output True labels, randomized labels

Diagnostic classifier

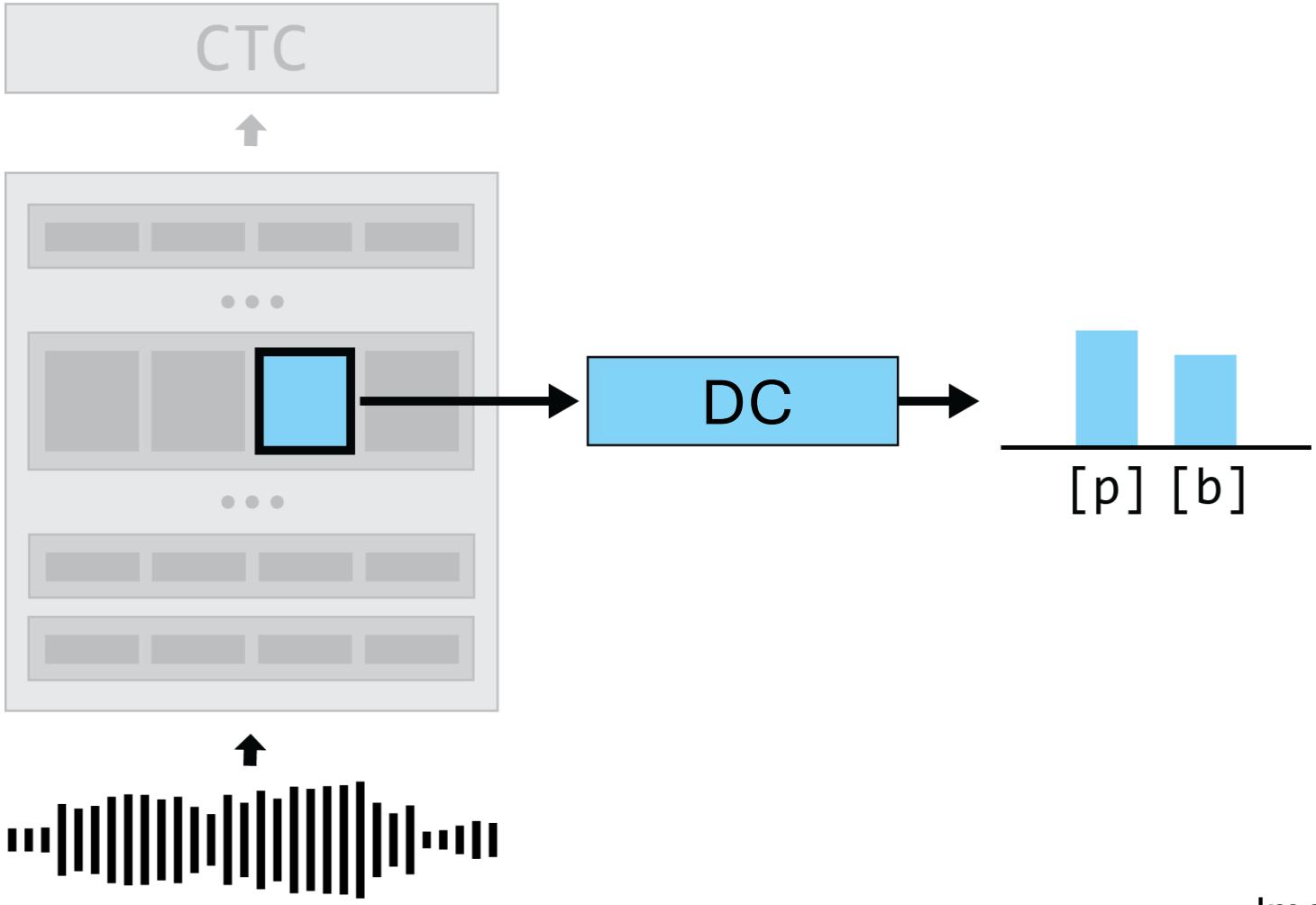
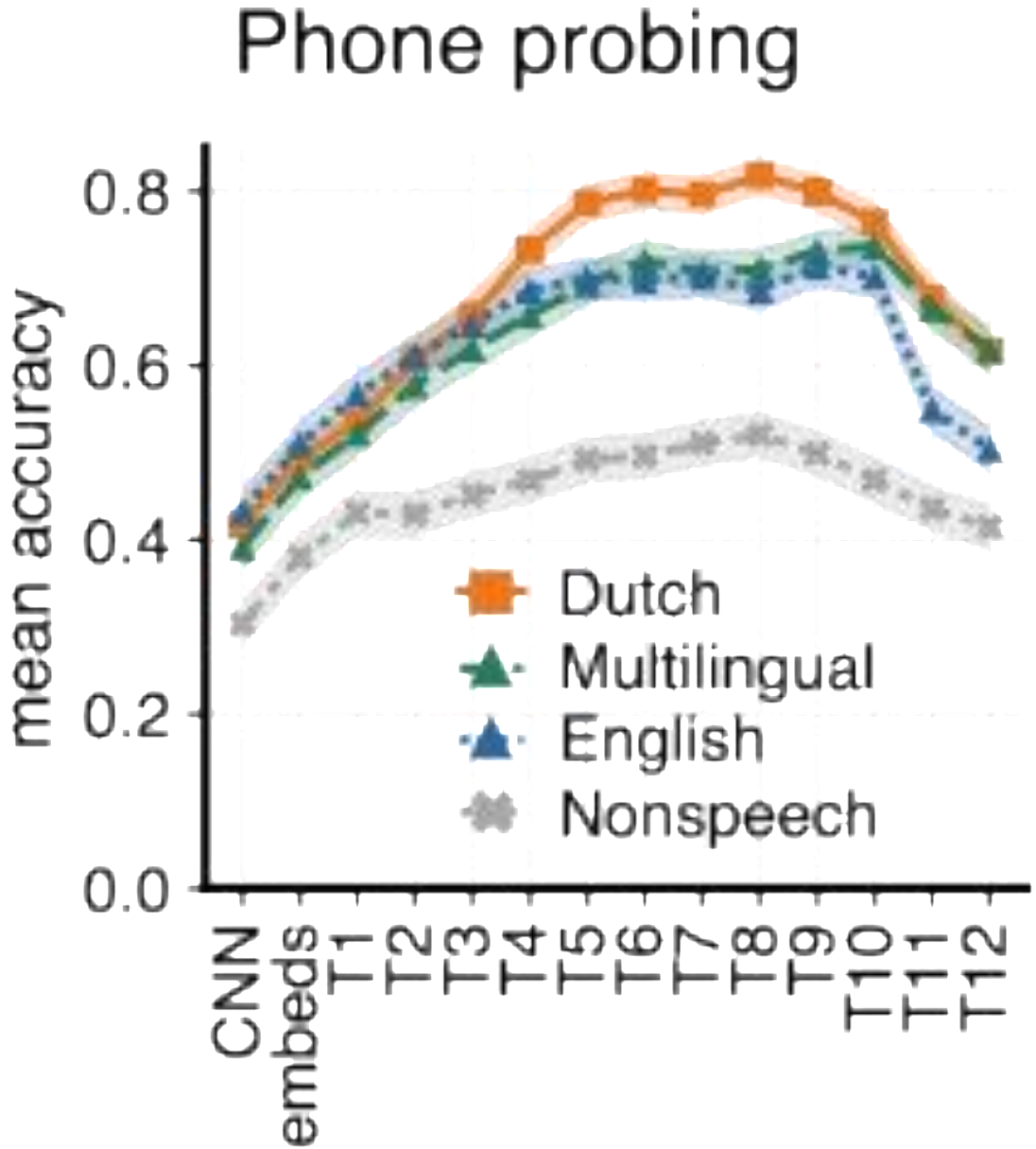


Image by Marianne de Heer Kloots



de Heer Kloots, M., Mohebbi, H., Pouw, C., Shen, G., Zuidema, W., Bentum, M. (2025) What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training. Proc. Interspeech 2025, 256-260

CTC lens

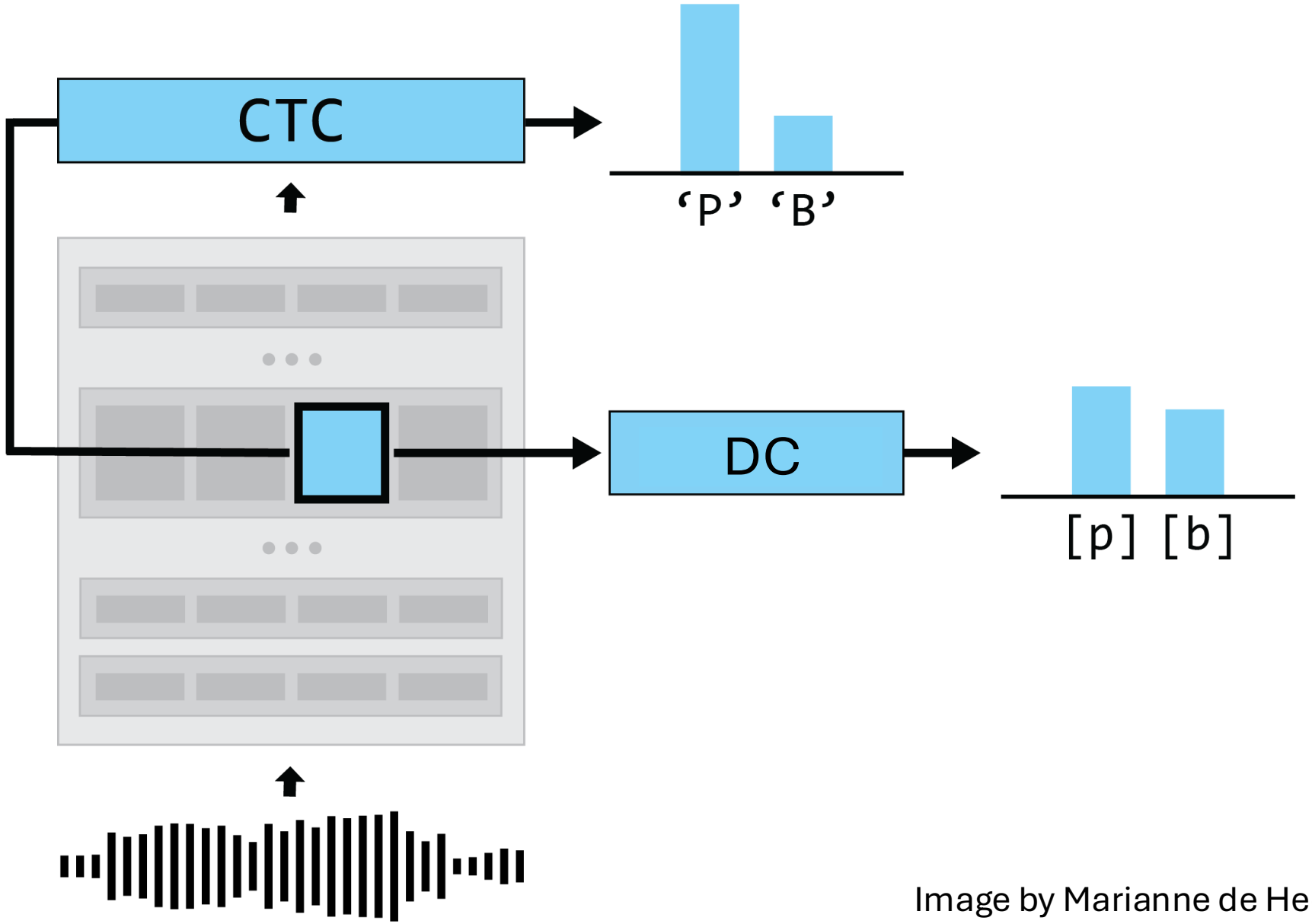


Image by Marianne de Heer Kloots

CTC lens

- Applying the model's own unembedding operations on earlier layers
- Wav2vec 2.0 CTC head on earlier transformer layers
- Compare with other probing techniques

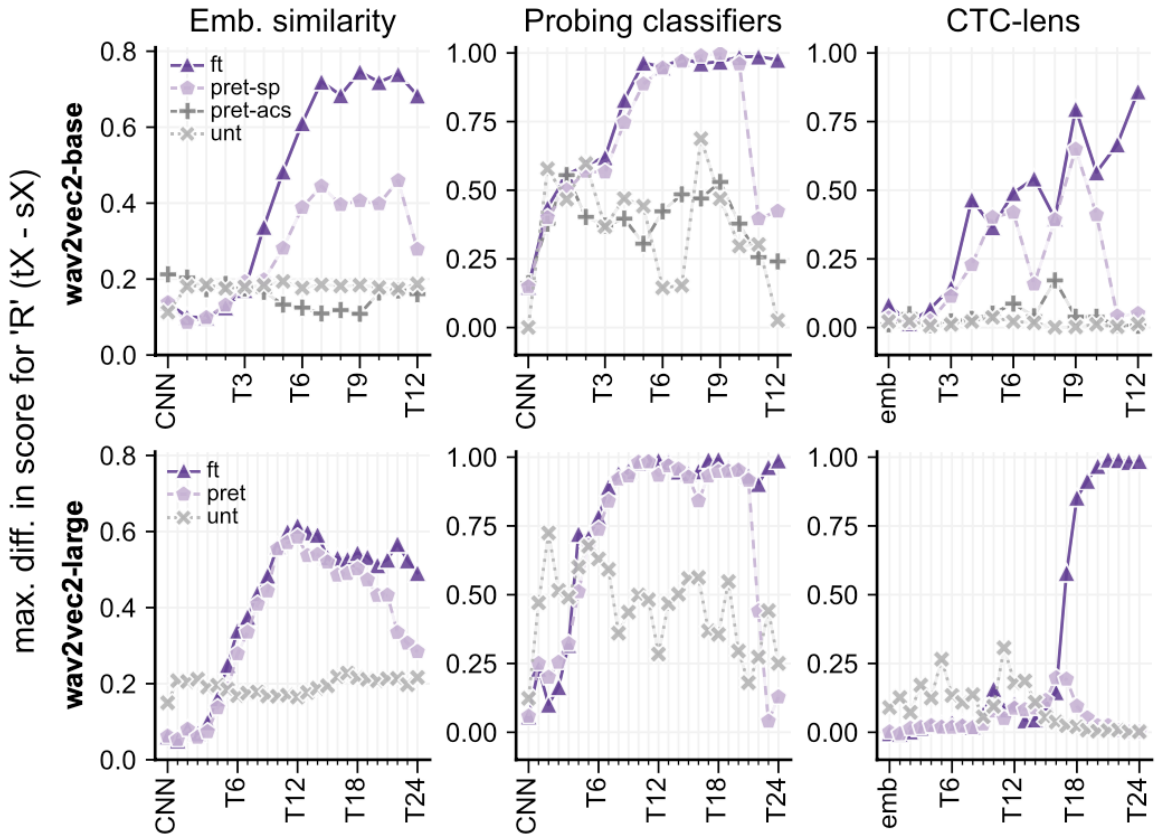
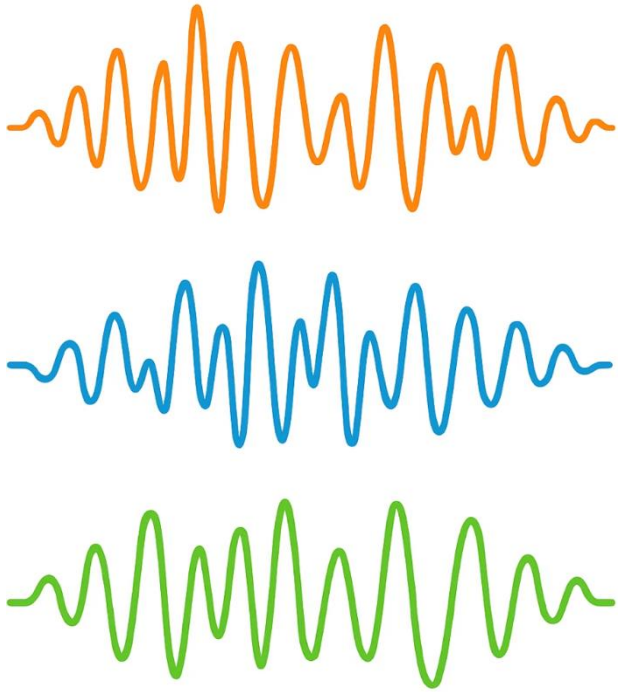
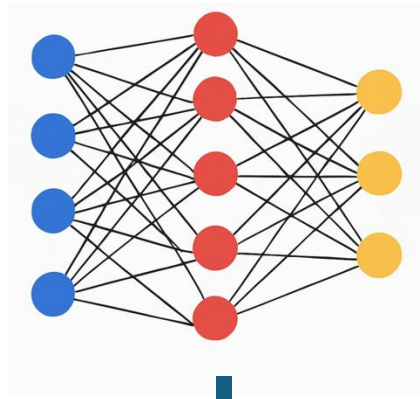


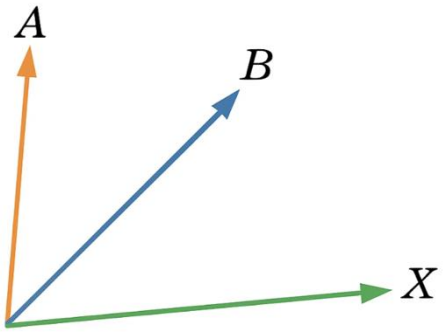
Figure 4: *Aggregated layerwise results for all analysis methods and two model sizes. Each point shows the highest difference in preference for 'R' between the tXih and sXih continua, across all intermediate continuum steps.*

de Heer Kloots, M., & Zuidema, W. (2024). Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2. 0. In *Proc. INTERSPEECH*.

ABX

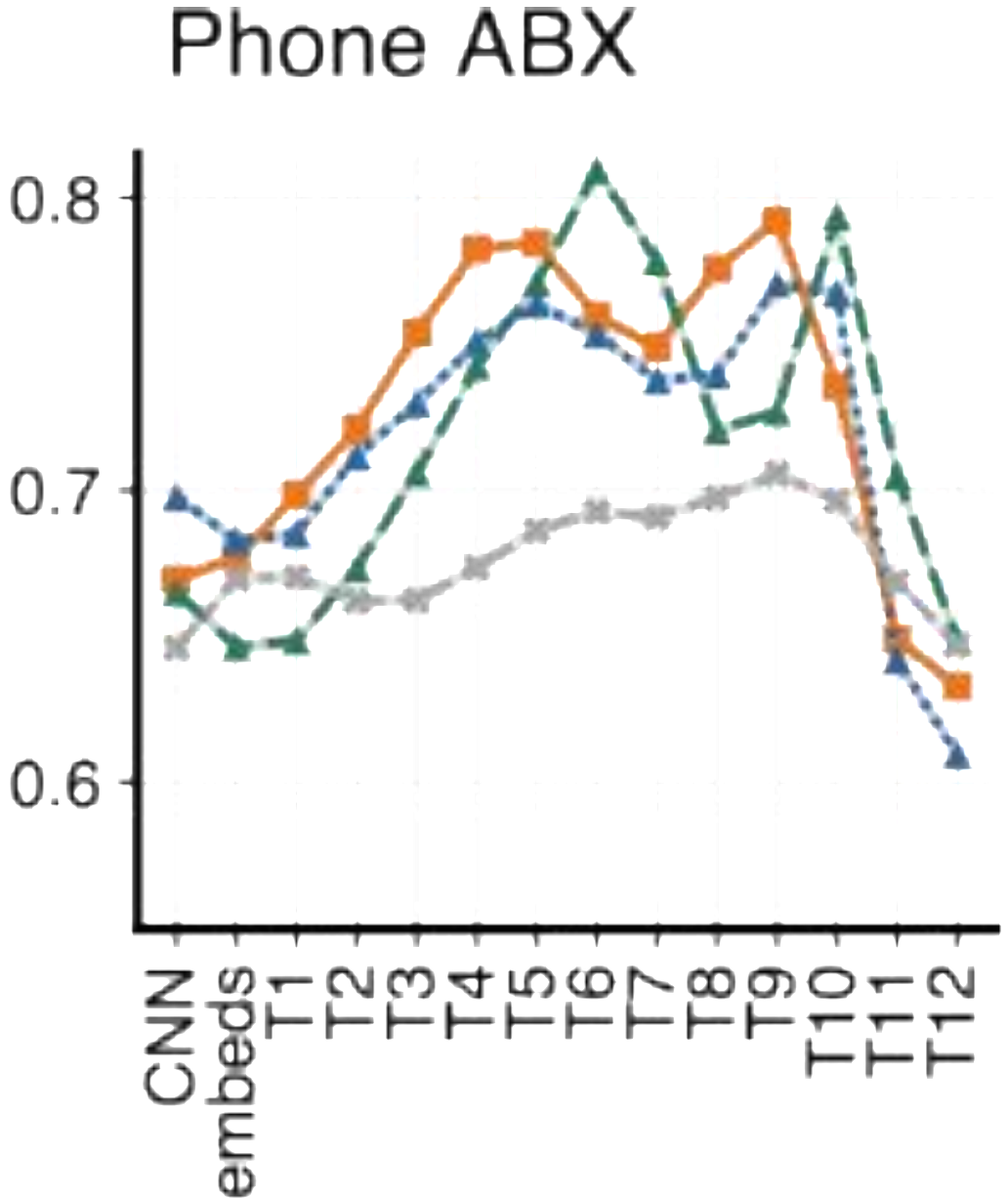






Is X similar to A or B ?

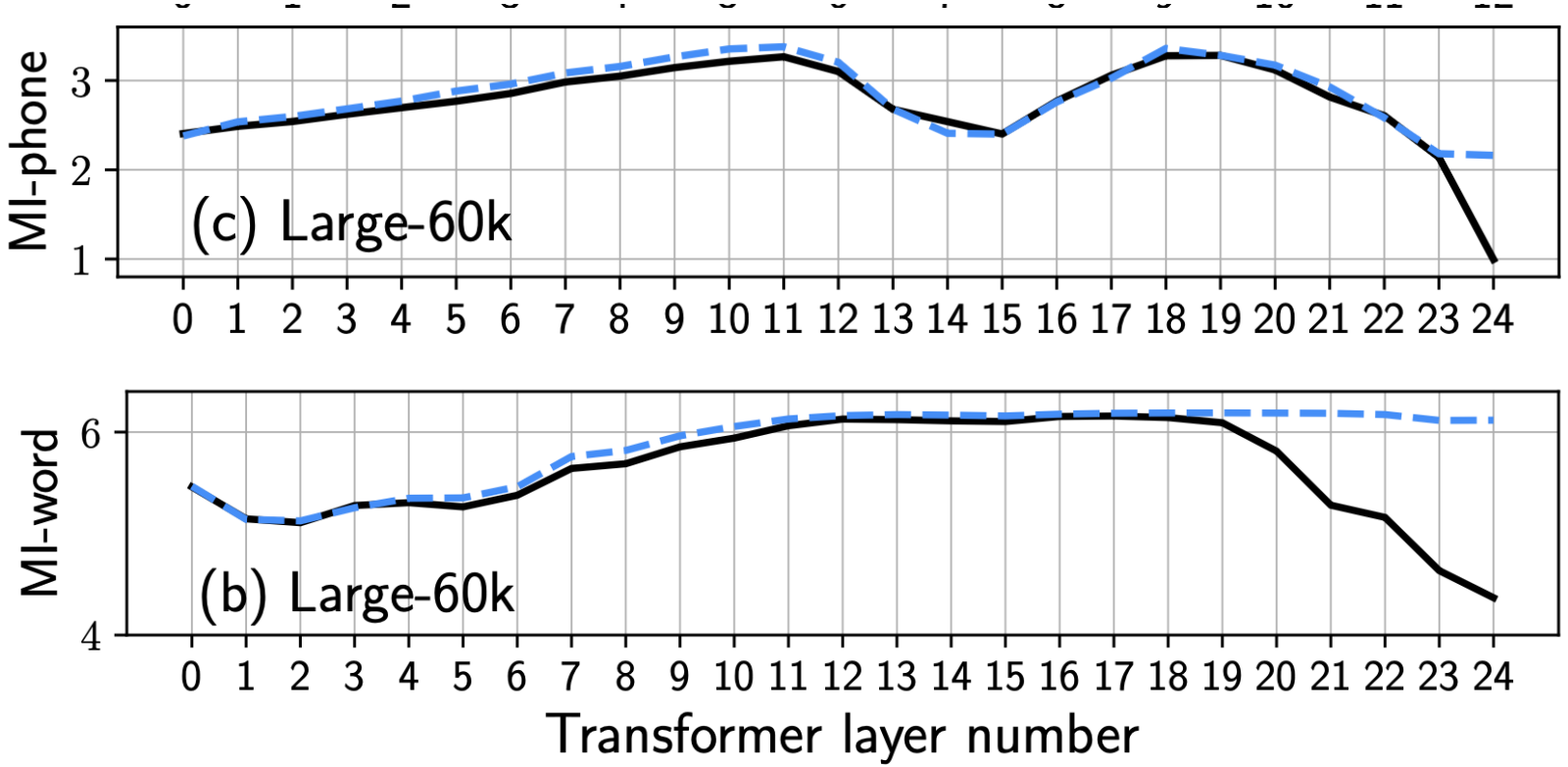
T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, “Evaluating speech features with the Minimal-Pair ABX task (I): Analysis of the classical MFC/PLP pipeline,” in *Proc. Interspeech*, 2013, pp. 1781–1785



de Heer Kloots, M., Mohebbi, H., Pouw, C., Shen, G., Zuidema, W., Bentum, M. (2025) What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training. Proc. Interspeech 2025, 256-260

Information theoretic measures

- Distributions of quantized model units can be compared to categorical linguistic units
- Wav2vec 2.0 quantizes CNN output to codevectors; HuBERT uses Kmeans clustering to map hidden states to discrete units
- **Mutual information**



Pasad, A., Chou, J. C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 914-921). IEEE.

Information theoretic measures

- Jensen-Shannon divergence

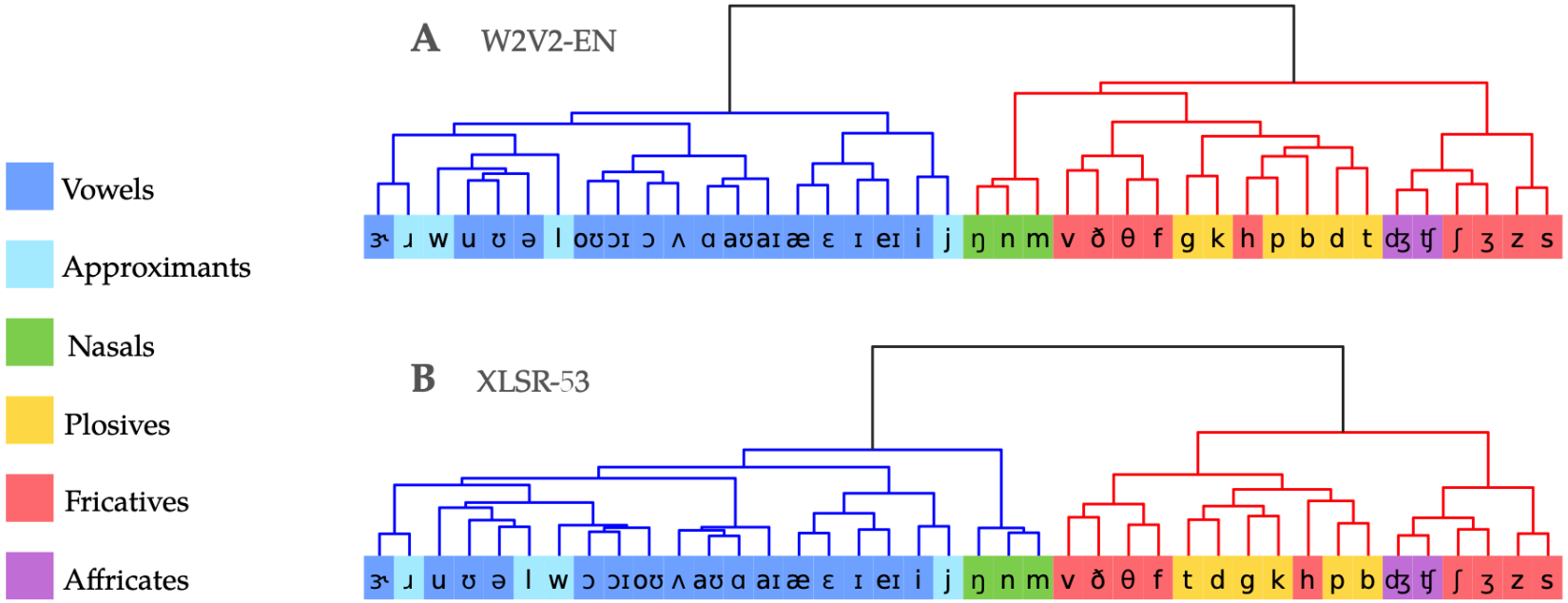
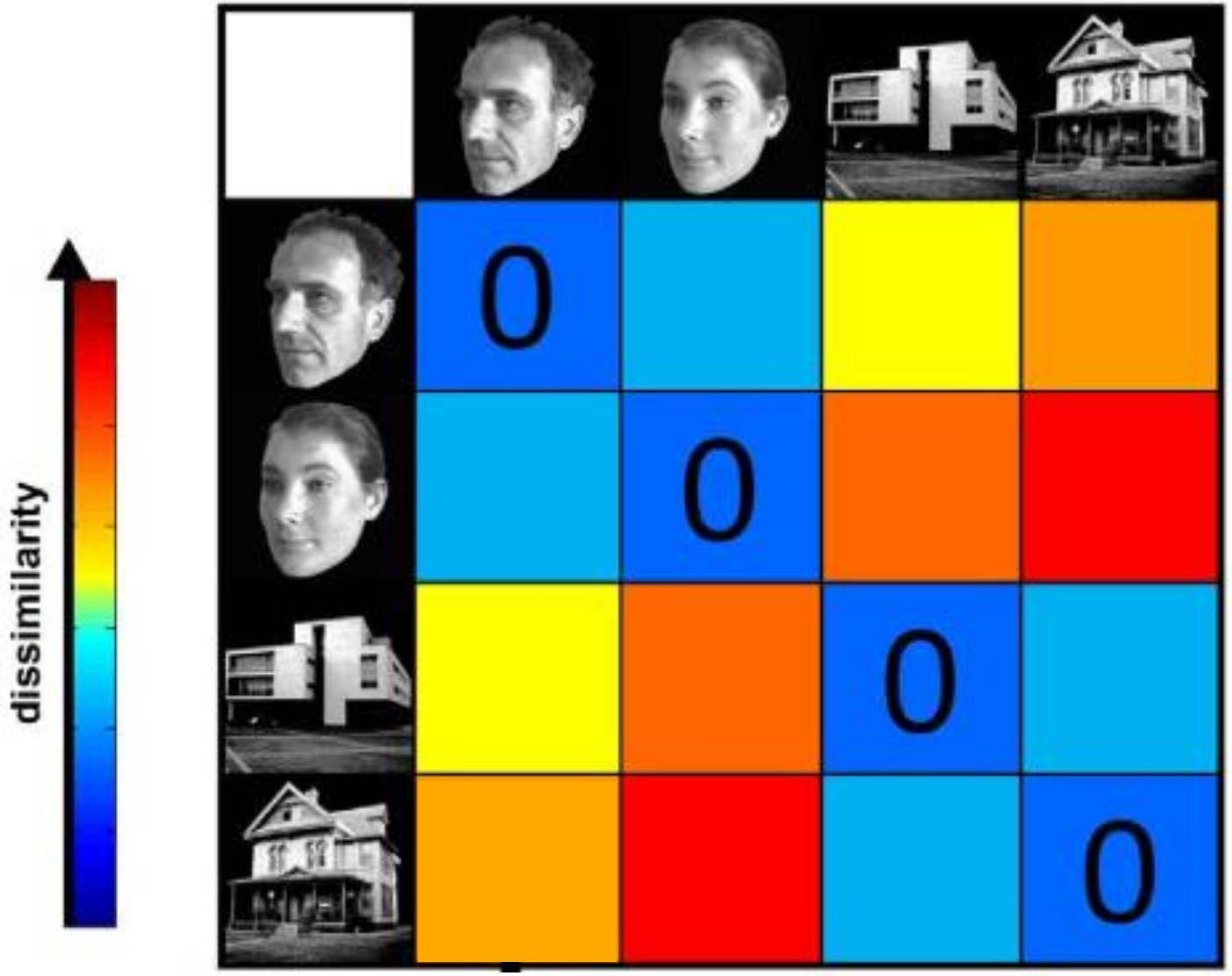
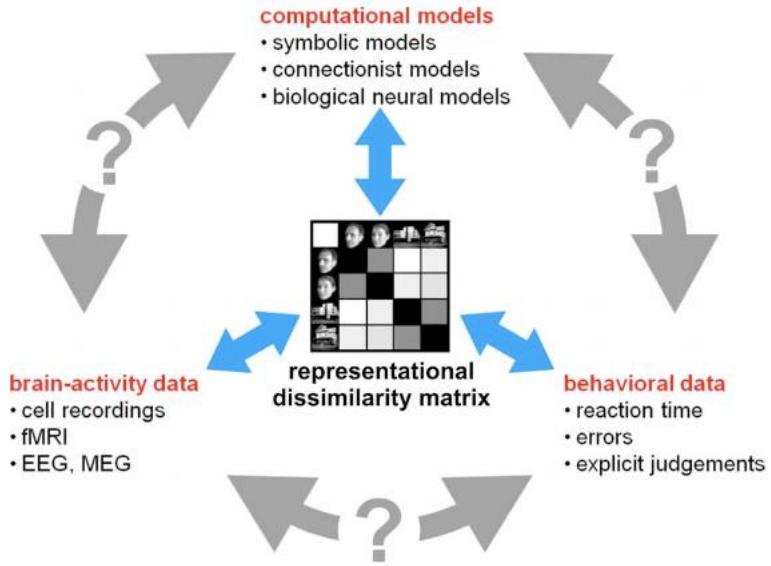


Figure 2: *The resulting clusters from applying agglomerative hierarchical clustering over the distance matrix, where our measure of the distance is the Jensen-Shannon divergence between phonetic distributions: (A)* W2V2 *and (B)* XLSR.

Abdullah, B.M., Shaik, M.M., Möbius, B., Klakow, D. (2023) An Information-Theoretic Analysis of Self-supervised Discrete Representations of Speech. Proc. Interspeech 2023, 2883-2887

Similarity spaces





Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

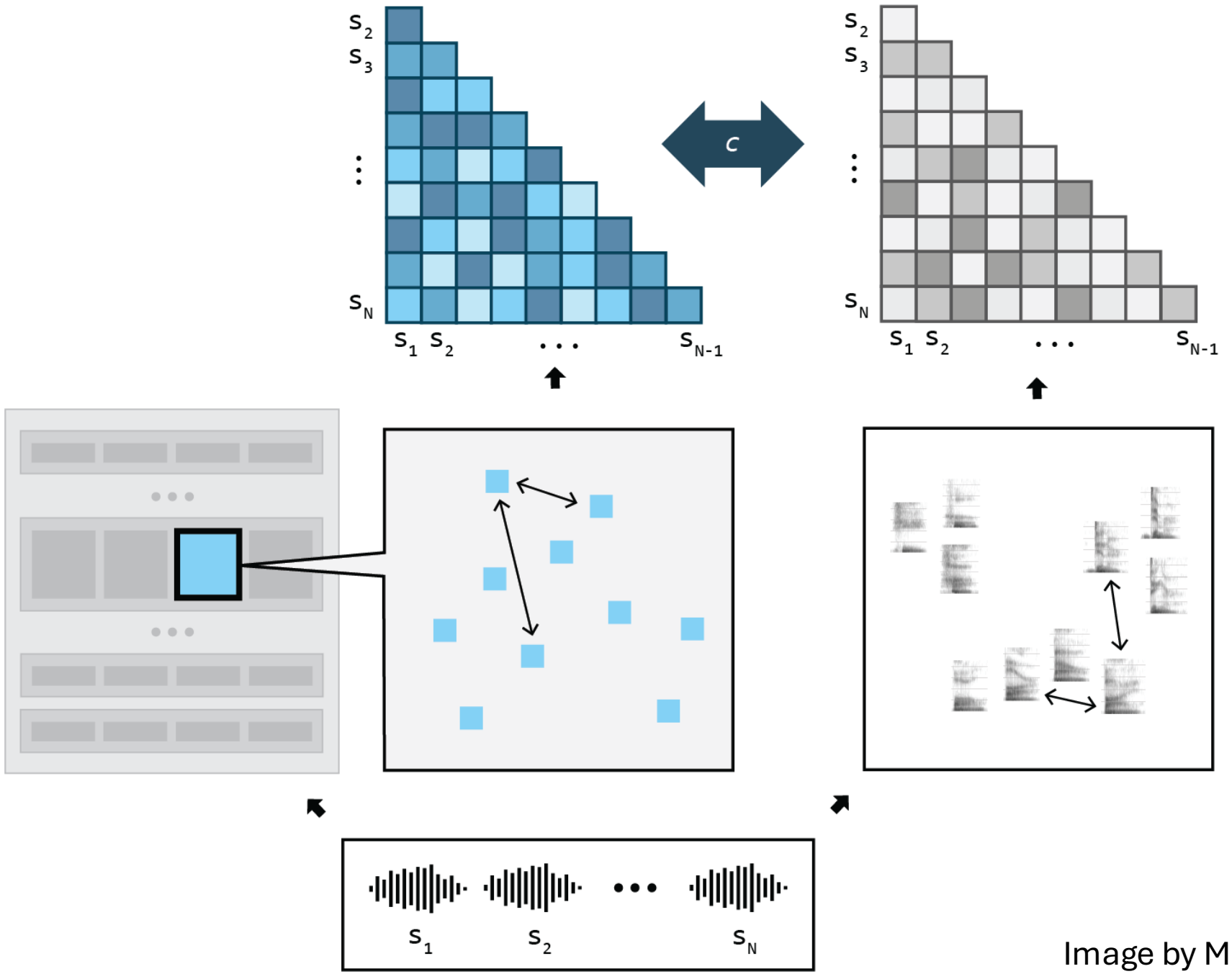
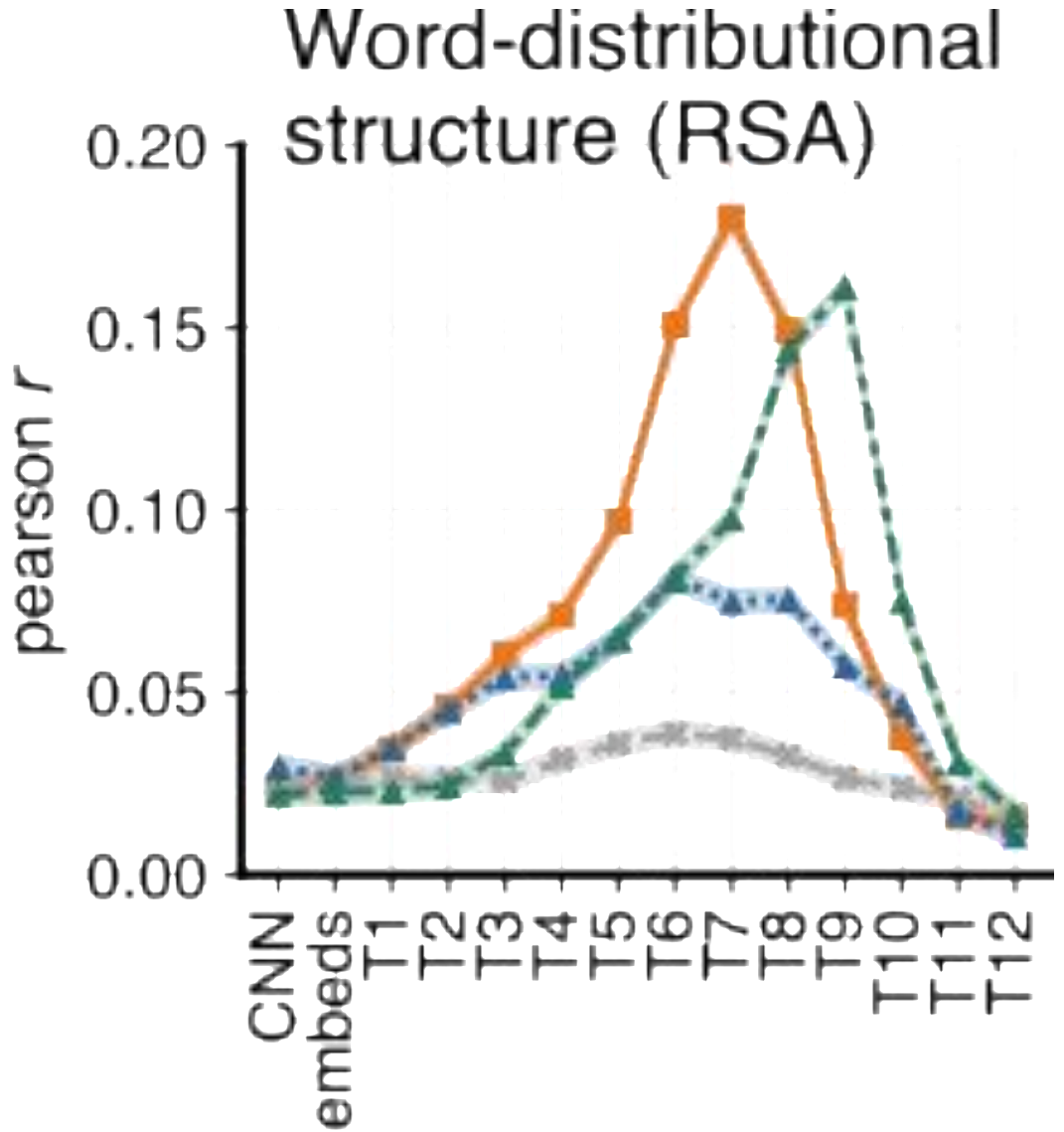


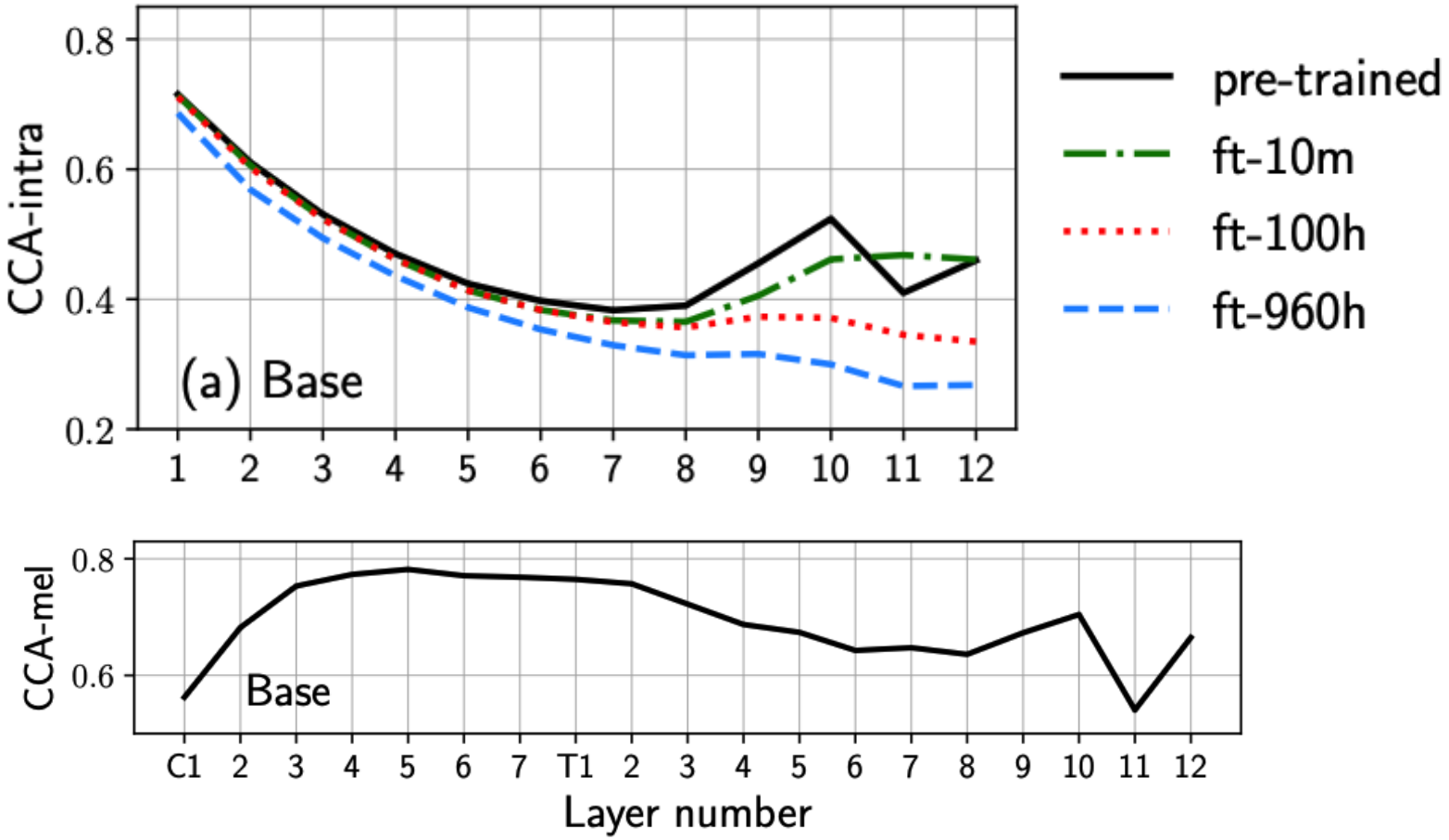
Image by Marianne de Heer Kloots



de Heer Kloots, M., Mohebbi, H., Pouw, C., Shen, G., Zuidema, W., Bentum, M. (2025) What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training. Proc. Interspeech 2025, 256-260

Similarity spaces

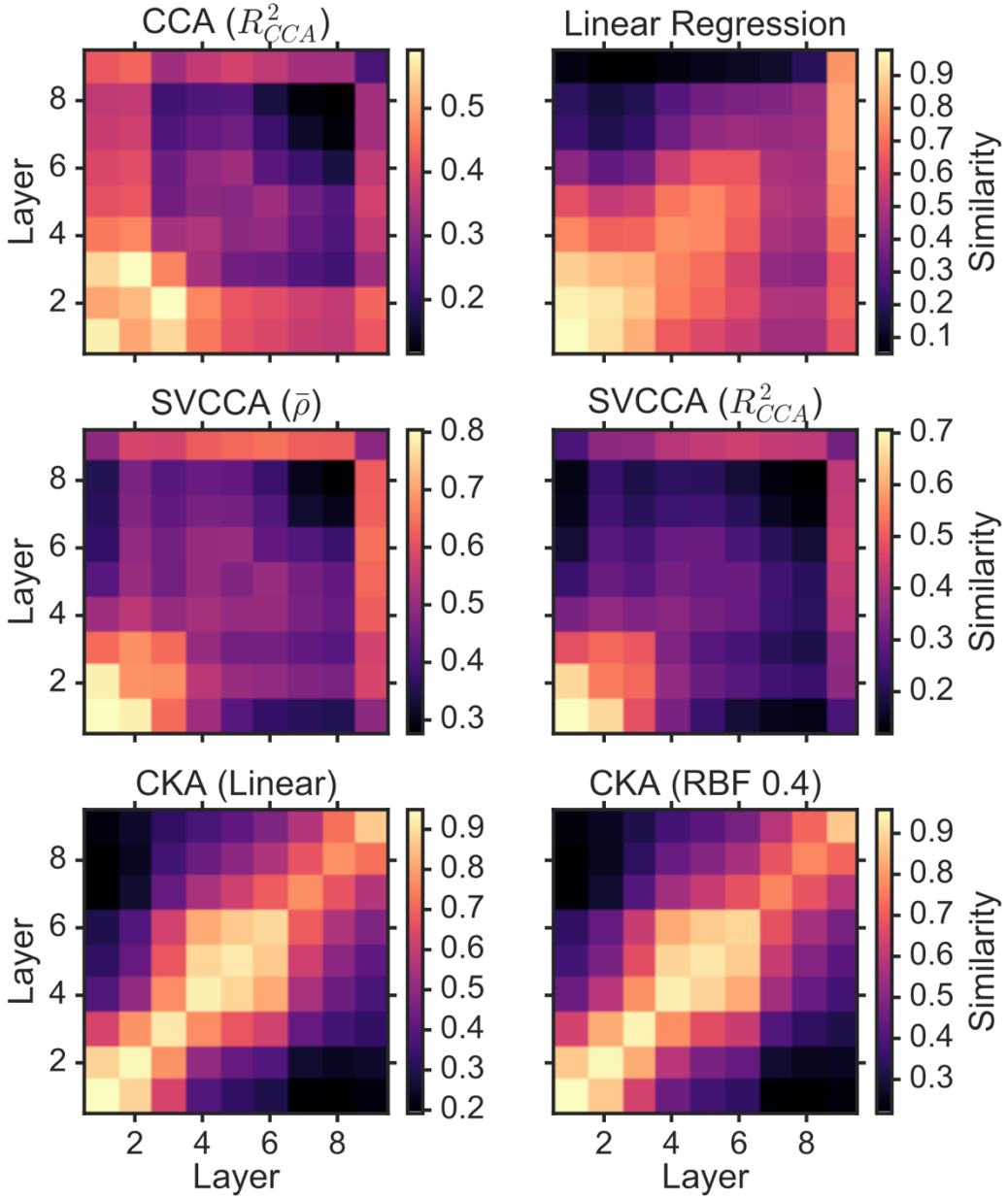
- CCA



Pasad, A., Chou, J. C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 914-921). IEEE.

Similarity spaces

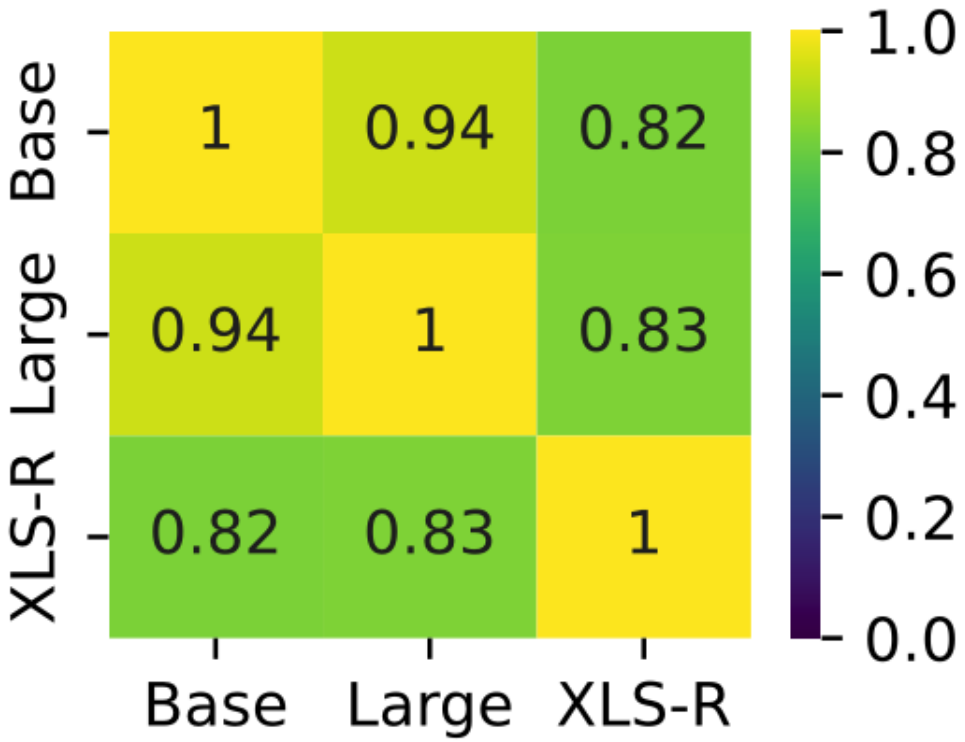
- CKA



Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, May). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519-3529). PMlR.

Similarity spaces

- CKA



Choi, K., & Yeo, E. J. (2022). Opening the black box of wav2vec feature encoder. *arXiv preprint arXiv:2210.15386*.

Overview of techniques

- Token based
  - ABX
  - Diagnostic classifier
  - CTC lens
- Similarity spaces
  - RSA
  - CCA
  - CKA
- Dimension reduction
  - PCA
  - LDA
  - MDS
  - t-SNE
  - UMAP
- Information theoric based
  - Mutual information
  - Jensen-Shannon distance