General Artificial Intelligence (1)

SAIR-2-06: Decoding Speech Perception from Non-Invasive Brain Recording

Momiao Xiong

Society of Artificial Intelligence Research

Decoding speech perception from non-invasive brain recordings

github.com/facebookresearch/brainmagick. The code is provided under the CC-NC-BY 4.0 license.

Nature machine intelligence, 5 October 2023

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, Alexei Baevski et al. 2020

Code and models are available at https://github.com/pytorch/fairseq

Learning Transferable Visual Models From Natural Language Supervision

code and pre-trained model weights at https://github.com/OpenAI/CLIP.

Speech Translation with Foundation Models and Optimal Transport: UPC at IWSLT23

EFFICIENT DOMAIN ADAPTATION FOR SPEECH FOUNDATION MODELS

https://github.com/savoirfairelinux/num2words

CONTRASTIVE AUDIO-VISUAL MASKED AUTOENCODER

Code and pretrained models are at https://github.com/yuangongnd/cav-mae.

High-resolution image reconstruction with latent diffusion models from human brain activity. bioRxiv, 2023.

Natural scene reconstruction from fMRI signals using generative latent diffusion

Furkan Ozcelik and Rufin VanRullen 2023

- **Decoding speech from brain activity is a long-awaited goal in both healthcare and neuroscience**

  Invasive devices have recently led to major milestones in this regard: deep-learning algorithms trained on intracranial recordings can now start to decode elementary linguistic features such as letters, words and .

  extending this approach to natural speech and non-invasive brain recordings remains a major challenge

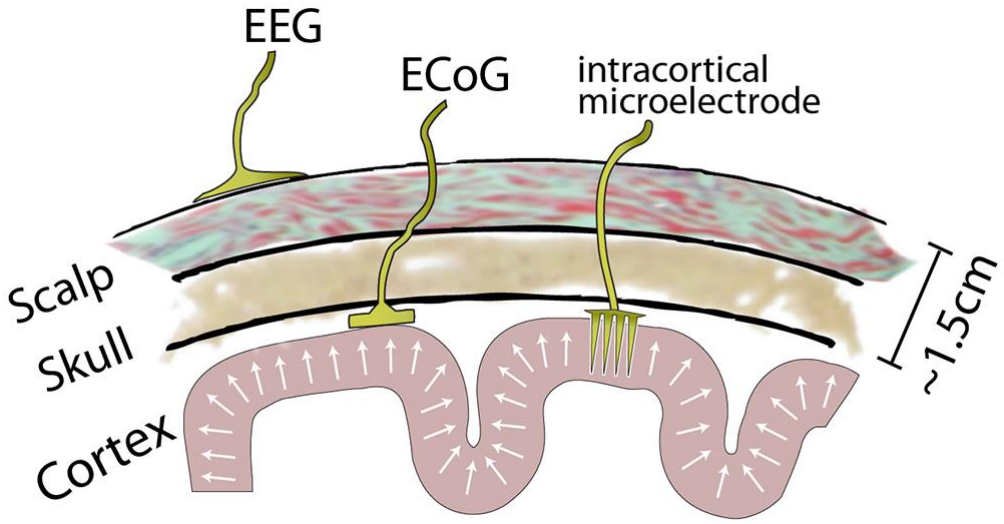
- **Here we introduce a model trained with contrastive learning to decode self-supervised representations of perceived speech from the non-invasive recordings of a large cohort of healthy individuals**

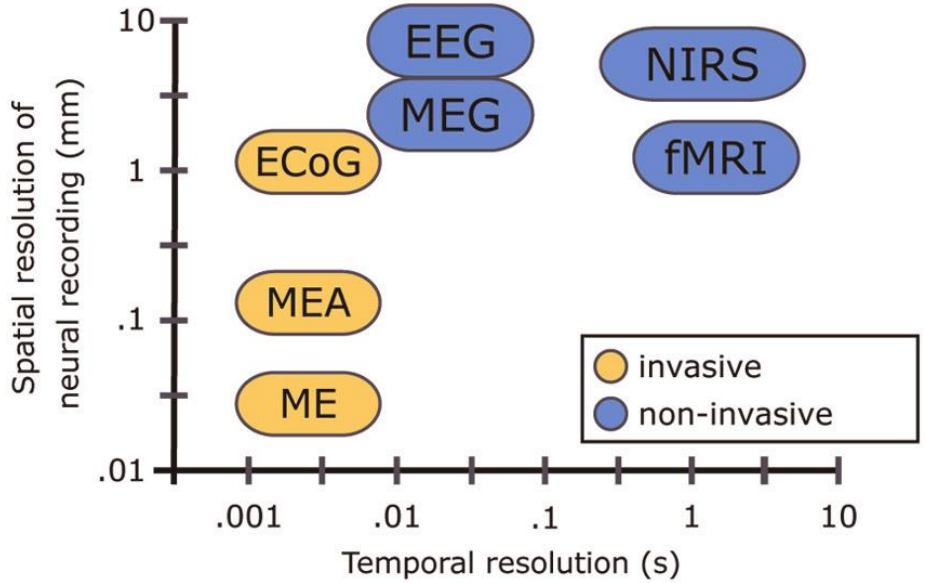
  curate and integrate four public datasets, encompassing 175 volunteers recorded with magneto-encephalography or electro-encephalography while they listened to short stories and isolated sentences.

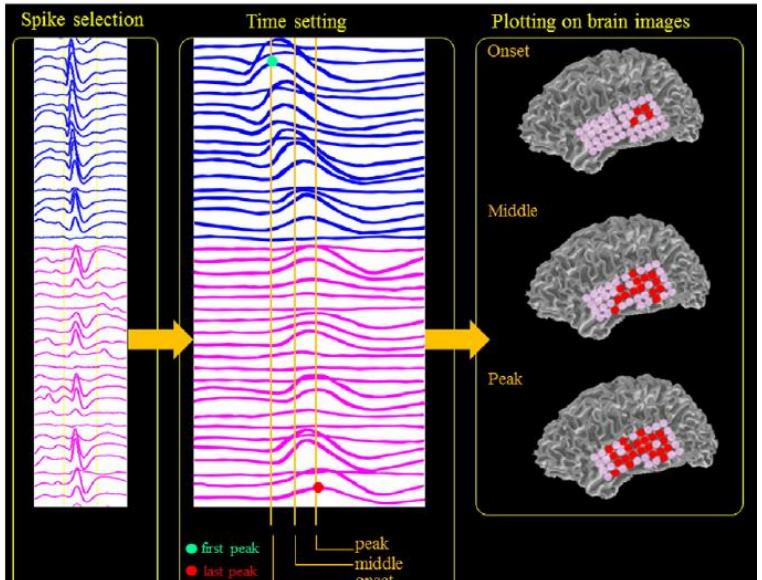
  allows the decoding of words and phrases absent from the training set.

  the analysis of the decoder's predictions suggests that they **primarily depend on lexical and contextual semantic representations**

- **Every year, traumatic brain injuries, strokes and neurodegenerative diseases cause thousands of patients lose their ability to speak or even communicate**







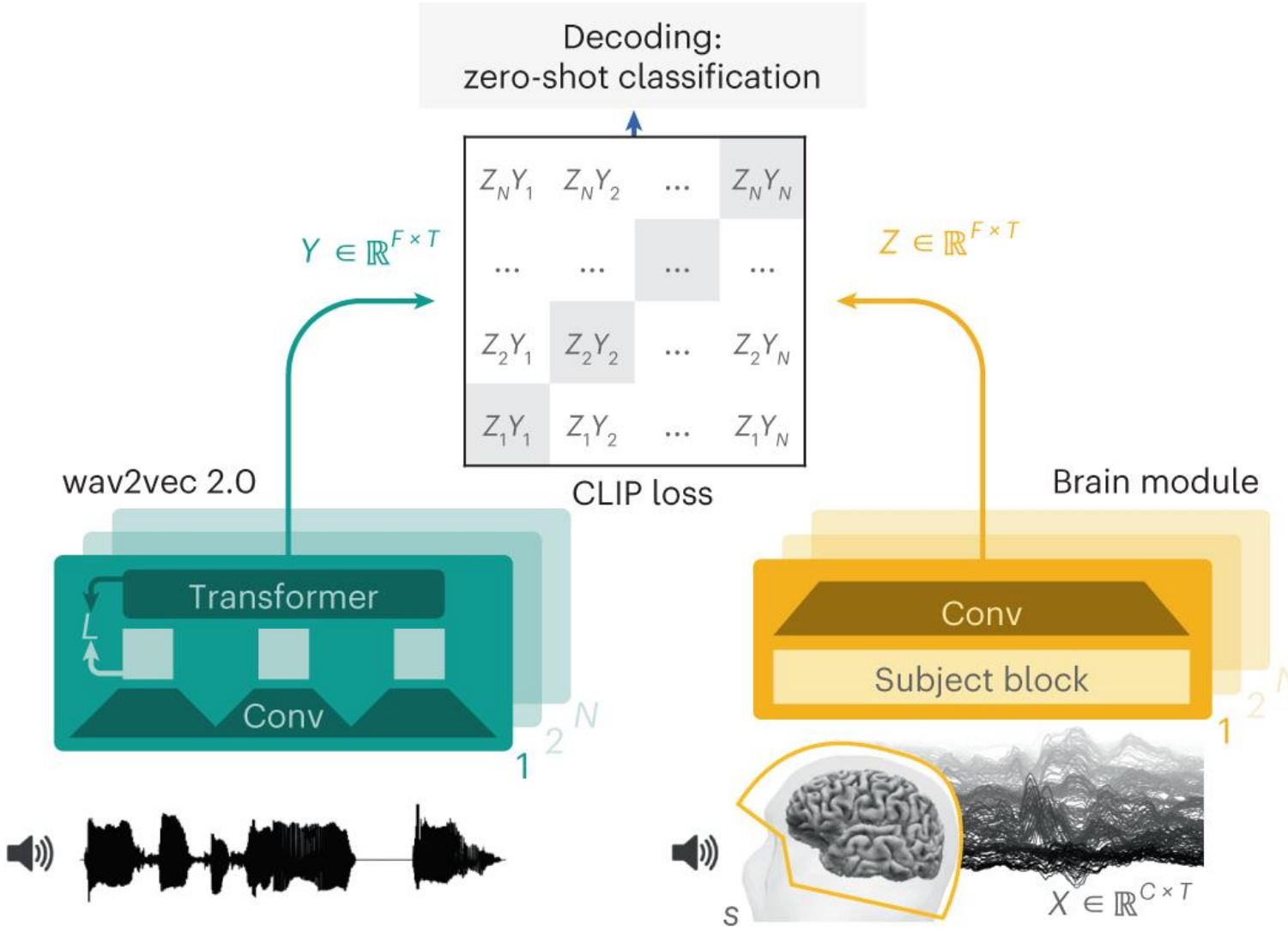
Proposed Approach

- **decode speech from non-invasive brain recordings by using**

  (1) a single architecture **trained** across a large cohort of participants

  (2) deep representations of speech learned with self-supervised learning on a large quantity of speech data.

- **focus the present work on speech perception in healthy volunteers rather than speech production in patients**



Model approach. We aim to decode speech from the brain activity of healthy participants recorded with MEG or EEG while they listen to stories and/ or sentences. For this, our model extracts the deep contextual representations of 3 s speech signals (Y of F feature by T time samples) from a pretrained speech module' (wav2vec 2.0: ref. 29).

learns the representations (Z) of the brain activity on the corresponding 3 s window (X of C recording channels by T time samples) that maximally align with these speech representations with a contrastive loss.

The representation Z is given by a deep convolutional network. At evaluation, we inpu t the model with left-out sentences and compute the probability of each 3 s speech segment given each brain representation.

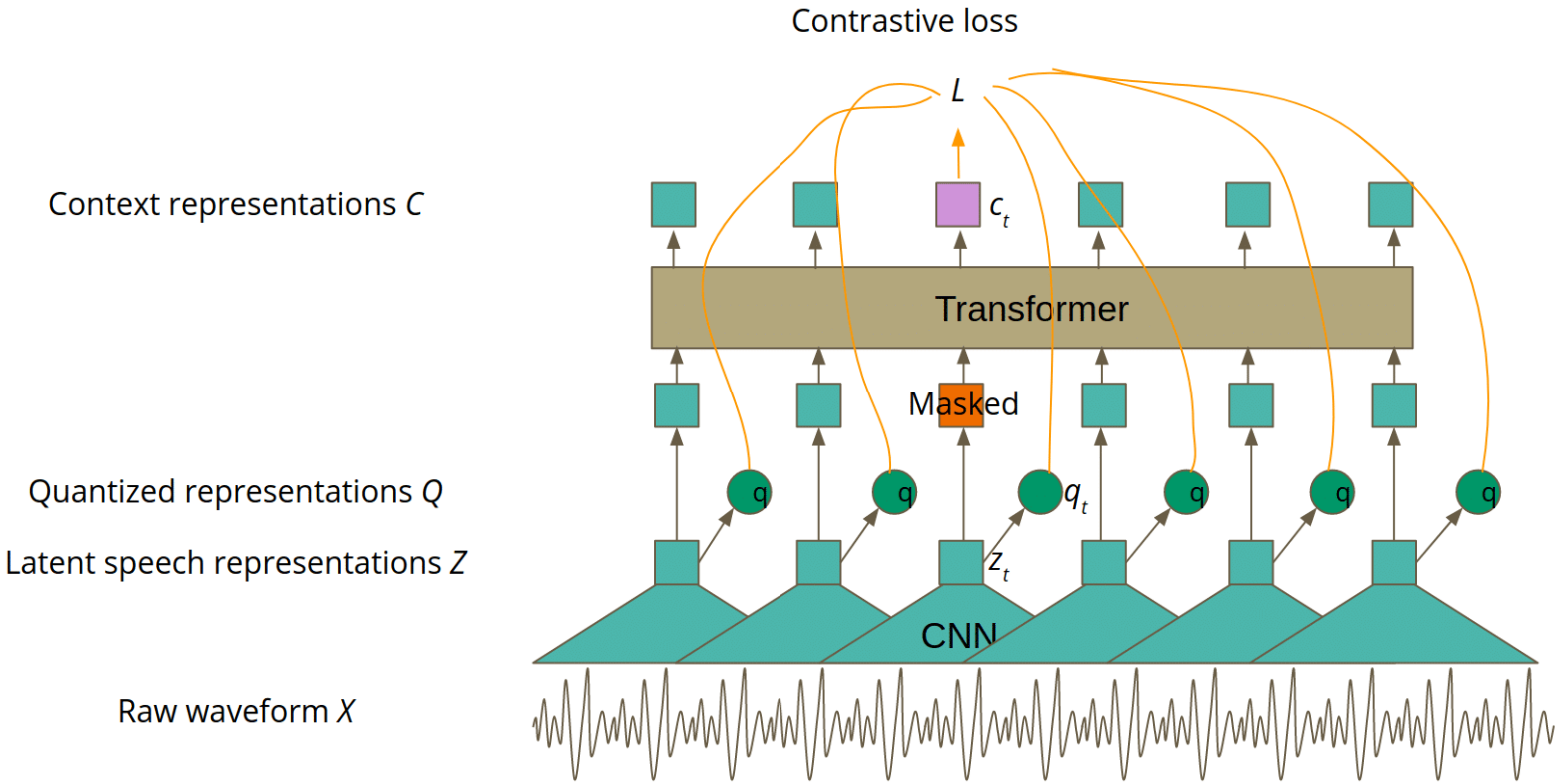
The resulting decoding can thus be 'zero shot' in that the audio snippets predicted by the model need not be present in the training set. This approach is thus more general than standard classification approaches where the decoder can only predict the categories learnt during training.

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

- encodes speech audio via a multi-layer convolutional neural network and then masks spans of the resulting latent speech representations
- The latent representations are fed to a Transformer network to build contextualized representations and the model is trained via a contrastive task where the true latent is to be distinguished from distractors

wav2vec 2.0

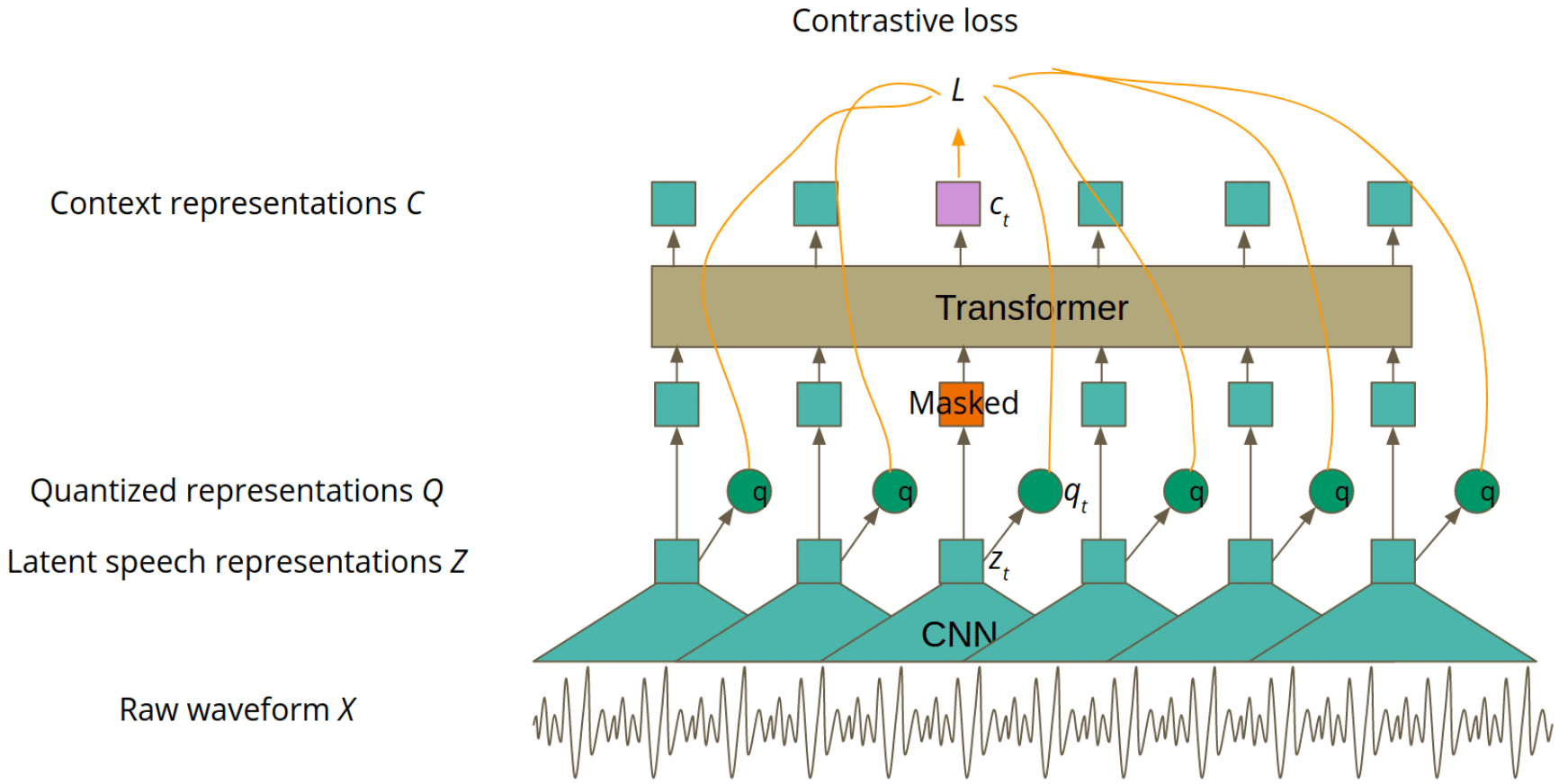
- encodes speech audio via a multi-layer convolutional neural network and then masks spans of the resulting latent speech representations
- The latent representations are fed to a Transformer network to build contextualized representations and the model is trained via a contrastive task where the true latent is to be distinguished from distractors



wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

An Illustrated Tour of Wav2vec 2.0

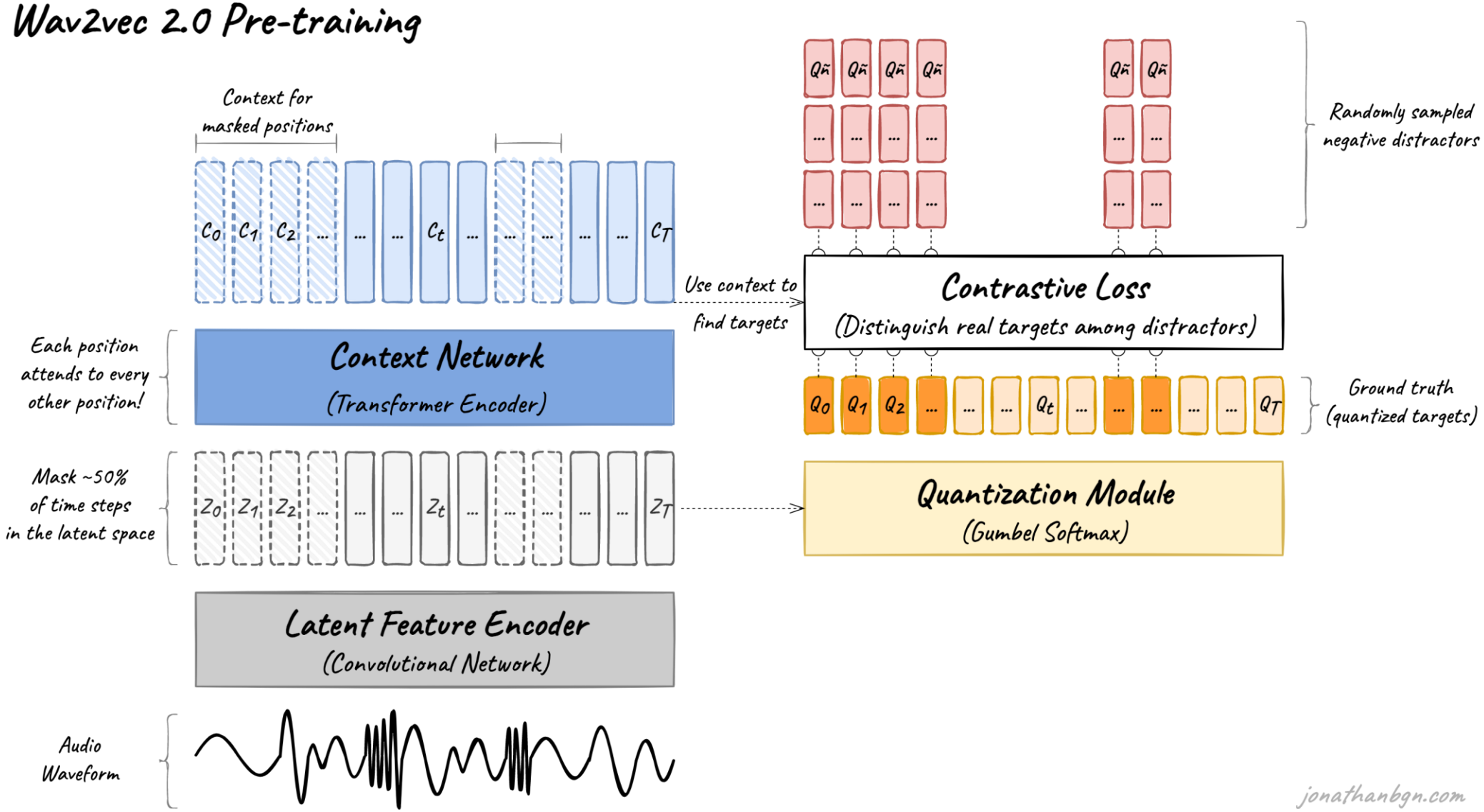
https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html



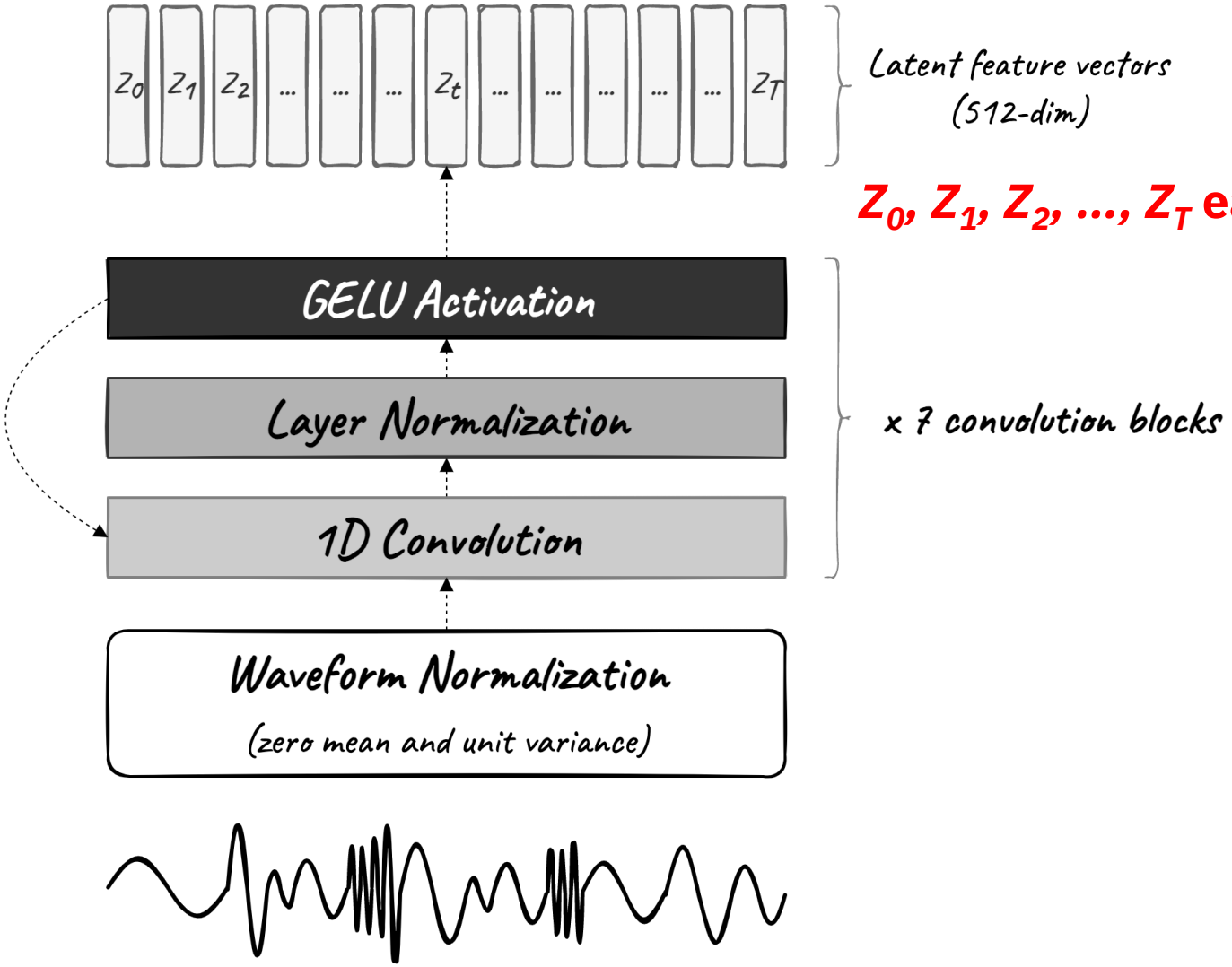
$$q_t = Q(z_t)$$

Quantization is a process of converting values from a continuous space into a finite set of values in a discrete space.

Wav2vec 2.0 is based on the Transformer's encoder, with a training objective similar to BERT's masked language modeling objective, but adapted for speech.



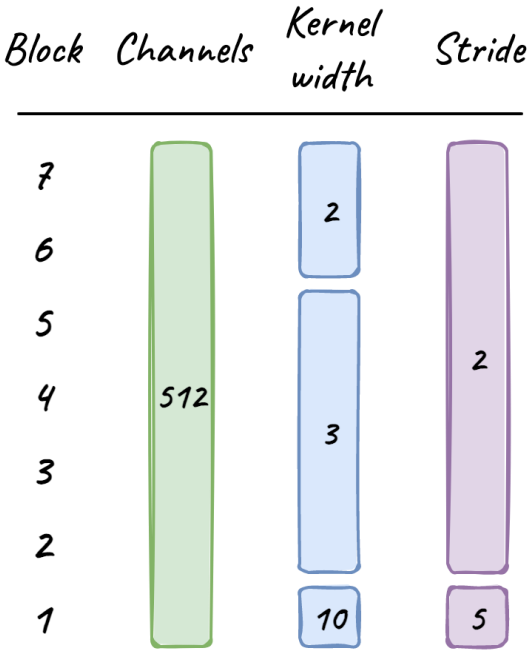
Wav2vec 2.0 Latent Feature Encoder



$$z_t = \begin{bmatrix} z_t^1 \\ \vdots \\ z_t^{512} \end{bmatrix}$$

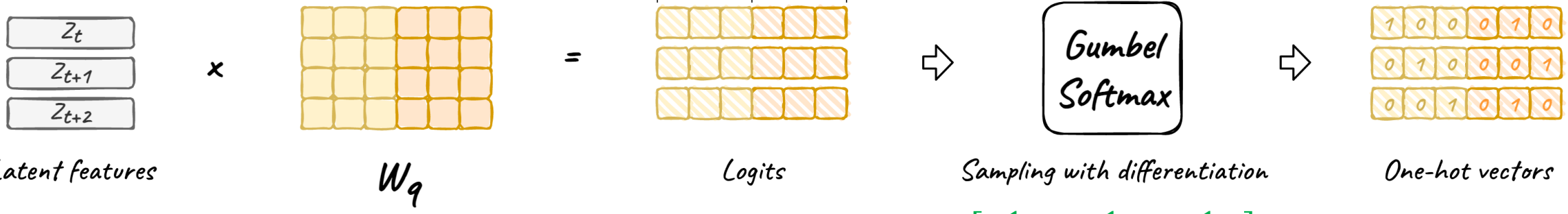
The feature encoder has a total receptive field of 400 samples or 25 ms of audio.

$z_0, z_1, z_2, \ldots, z_T$ each 20 milliseconds.



Wav2vec 2.0 Quantization Module

$$[z_t^1 \quad \cdots \quad z_t^{512}] \begin{bmatrix} w_{11}^1 & w_{12}^1 & w_{13}^1 \\ \vdots & \vdots & \vdots \\ w_{512,1}^1 & w_{512,2}^1 & w_{512,3}^1 \end{bmatrix} = \begin{bmatrix} e_{11}^1 & e_{12}^1 & e_{13}^1 \\ \vdots & \vdots & \vdots \\ e_{512,1}^1 & e_{512,2}^1 & e_{512,1}^1 \end{bmatrix}$$



Latent features

W_q

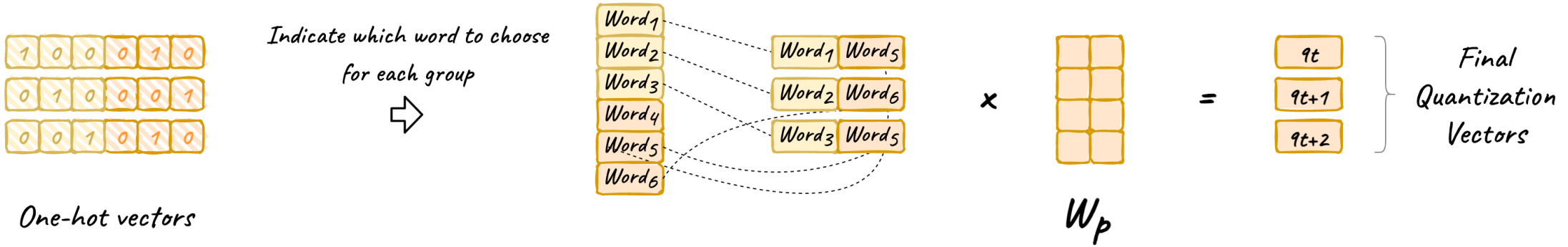
Quantization Matrix

Logits

Sampling with differentiation

One-hot vectors

$$l^1 = \begin{bmatrix} l_{11}^1 & l_{12}^1 & l_{13}^1 \\ \vdots & \vdots & \vdots \\ l_{512,1}^1 & l_{512,2}^1 & l_{512,3}^1 \end{bmatrix}, l_{i,j} = \log e_{ij}^1$$



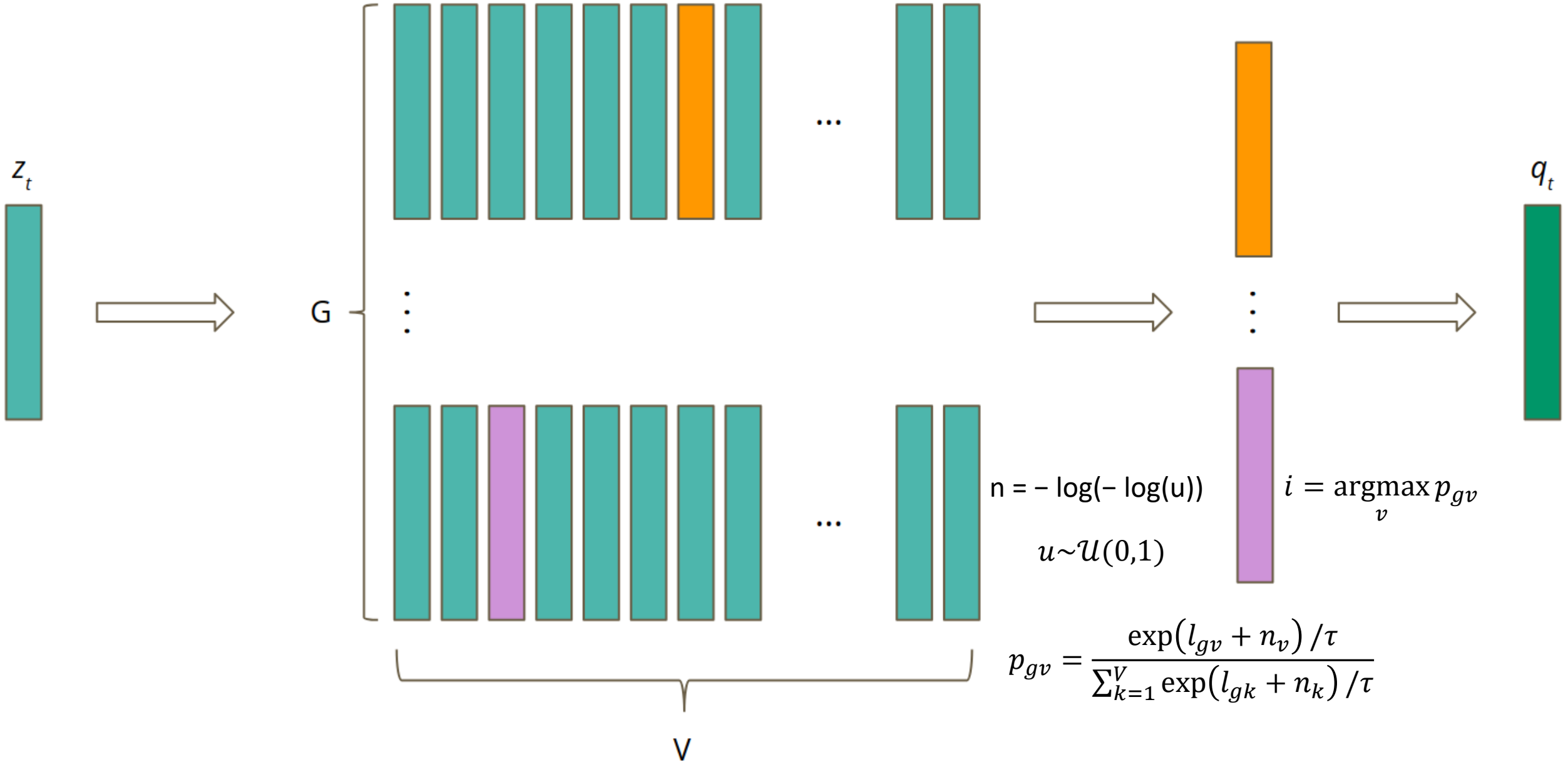
One-hot vectors

W_p

Quantization Projection Matrix

$$[z_t^1 \quad \cdots \quad z_t^{512}] \begin{bmatrix} w_{11}^2 & w_{12}^2 & w_{13}^2 \\ \vdots & \vdots & \vdots \\ w_{512,1}^2 & w_{512,2}^2 & w_{512,3}^2 \end{bmatrix} = \begin{bmatrix} e_{11}^2 & e_{12}^2 & e_{13}^2 \\ \vdots & \vdots & \vdots \\ e_{512,1}^2 & e_{512,2}^2 & e_{512,1}^2 \end{bmatrix}$$

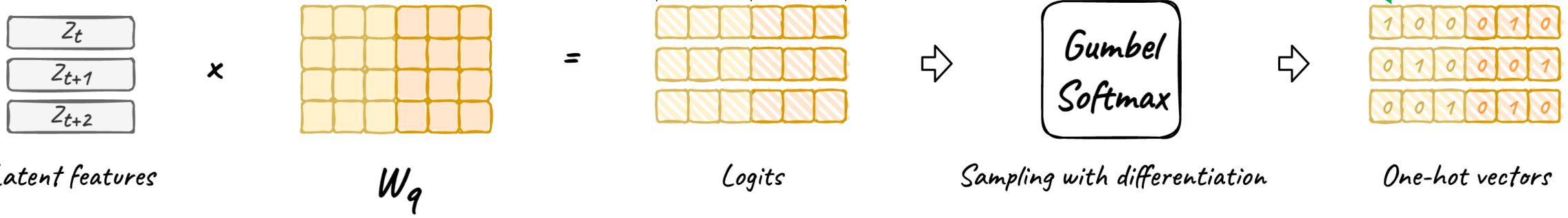
$$l^2 = \left(l_{ij}^2 = \log e_{ij}^2\right)$$


z_t
G
V
q_t
n = – log(– log(u))
$u \sim \mathcal{U}(0,1)$
$i = \underset{v}{\operatorname{argmax}}\, p_{gv}$
$p_{gv} = \frac{\exp\left(l_{gv} + n_v\right)/\tau}{\sum_{k=1}^{V} \exp\left(l_{gk} + n_k\right)/\tau}$

Wav2vec 2.0 Quantization Module

$$p_{gv} = \frac{\exp\left(l_{gv} + n_v\right)/\tau}{\sum_{k=1}^{V} \exp\left(l_{gk} + n_k\right)/\tau}$$

$$i = \underset{v}{\operatorname{argmax}}\, p_{gv}$$



Latent features

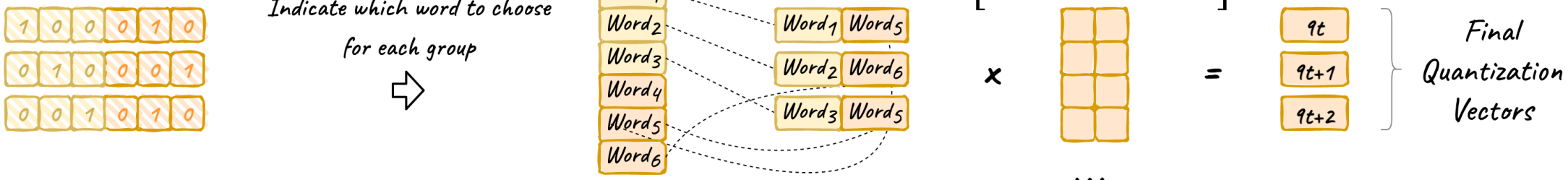
W_q

Quantization Matrix

Logits

Sampling with differentiation

One-hot vectors



One-hot vectors

Indicate which word to choose for each group

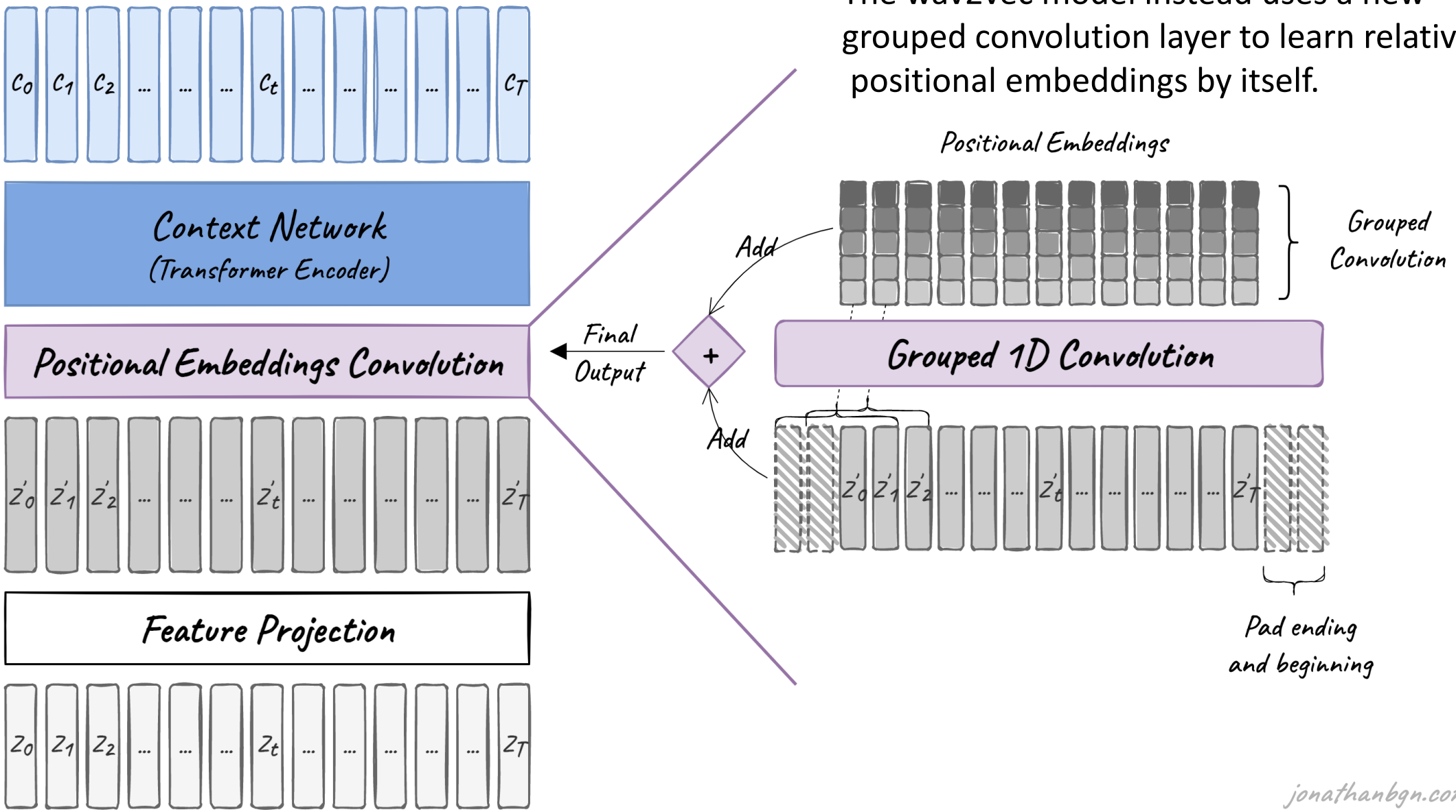
W_p

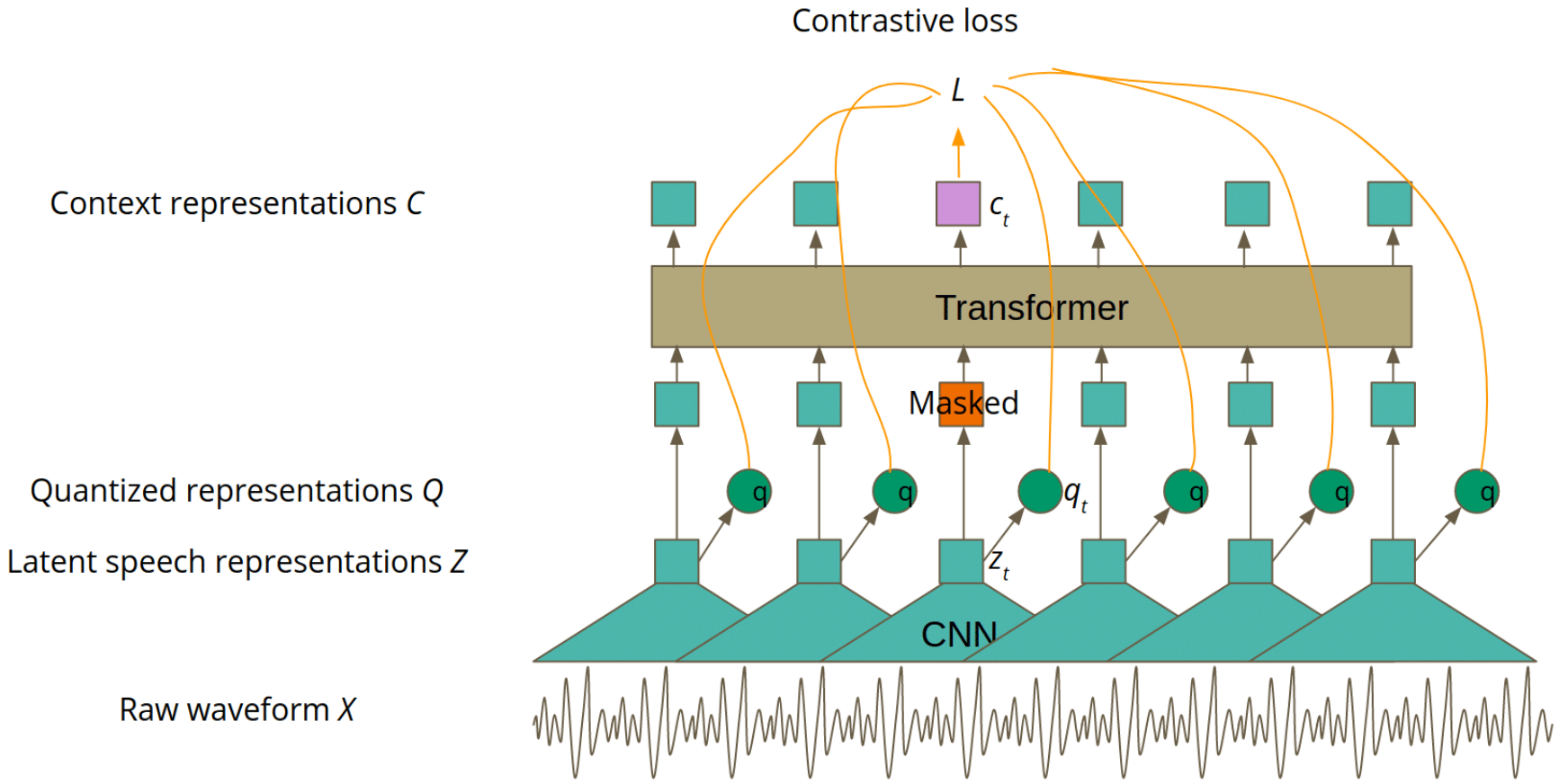
Quantization Projection Matrix

Final Quantization Vectors

Wav2vec 2.0 Context Network (Transformer Encoder)

The wav2vec model instead uses a new grouped convolution layer to learn relative positional embeddings by itself.



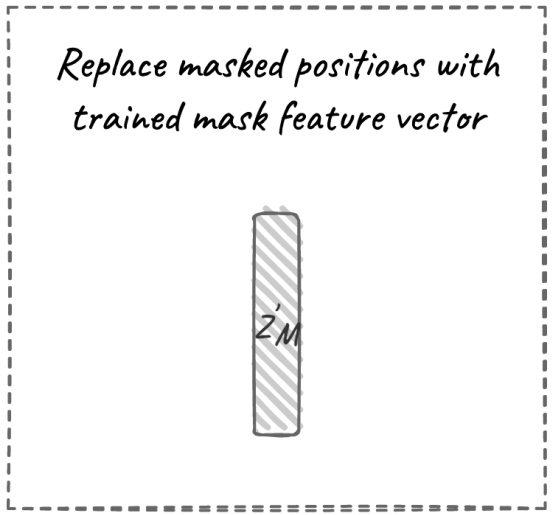


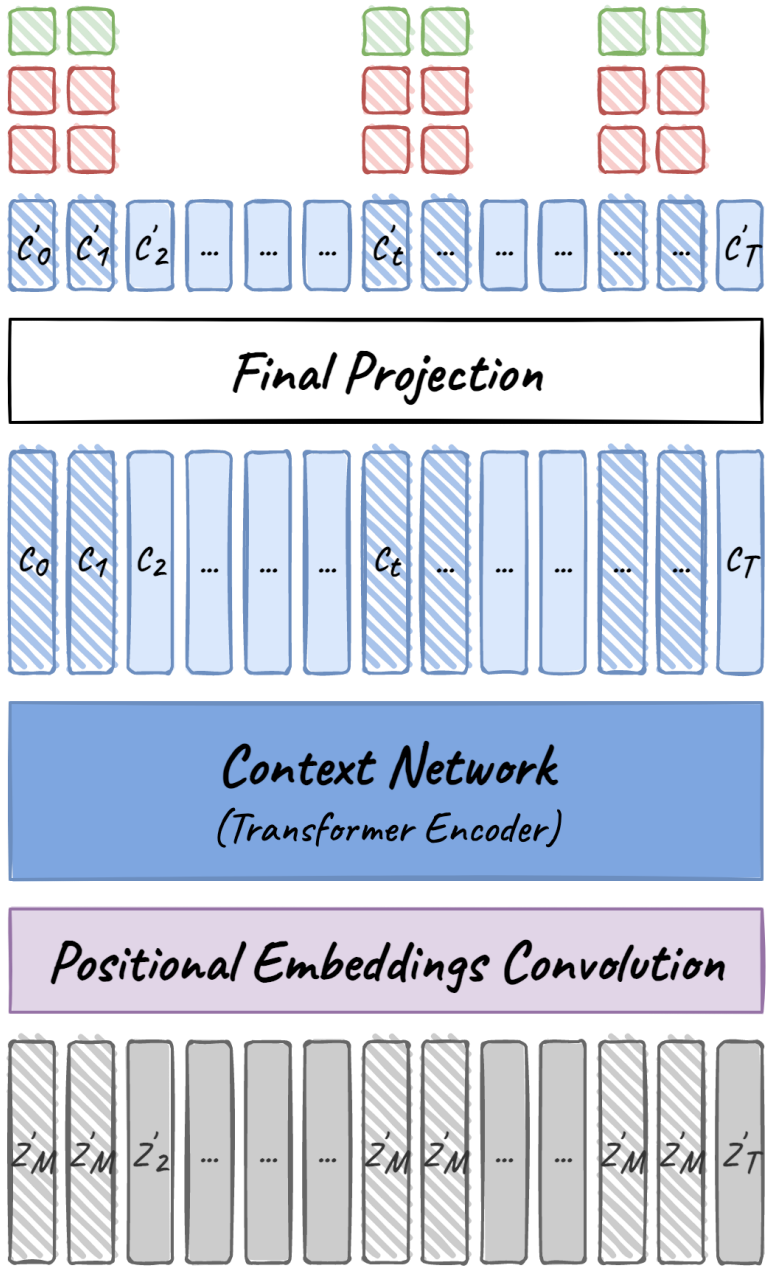
$$q_t = Q(z_t)$$

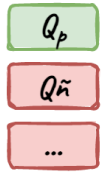
Quantization is a process of converting values from a continuous space into a finite set of values in a discrete space.

Wav2vec 2.0 Contrastive Loss

For each masked position, 100 negative distractors are uniformly sampled from other positions in the same sentence.







Compute similarity between final context vector $\hat{C}_i$ and positive / negative targets

The pre-training process uses a contrastive task to train on unlabeled speech data. A mask is first randomly applied in the latent space, where ~50% of the projected latent feature vectors. Masked positions are then replaced by the same trained vector Z'M before being fed to the Transformer network.

Randomly mask ~50% of the projected latent feature vectors Z'i

Training

- **Objective**
- **Contrastive Loss**

Given context network output C_t centered over masked time step t, the model needs to identify the true quantized latent speech representation q_t,

in a set of K + 1 quantized candidate representations $\tilde{q} \in Q_t$ which includes q_t and K distractors.

Distractors are uniformly sampled from other masked time steps of the same utterance

$$\mathcal{L}_m = -\log \frac{exp\big(sim(c_t', q_t)/\kappa\big)}{\sum_{\tilde{q}\in Q_t} \exp(sim(c_t', \tilde{q})/\kappa)}.$$

- **Diversity Loss**

The contrastive task depends on the codebook to represent both positive and negative examples and the diversity loss Ld is designed to increase the use of the quantized codebook representations. We encourage the equal use of the V entries in each of the G codebooks by maximizing the entropy of the averaged softmax distribution over the codebook entries for each codebook $\bar{p}_g$ across a batch of utterances; the softmax disribution does not contain the gumbel noise nor a temperature.

$$\mathcal{L}_d = \frac{1}{G}\sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} p_{gv}\log p_{gv}$$

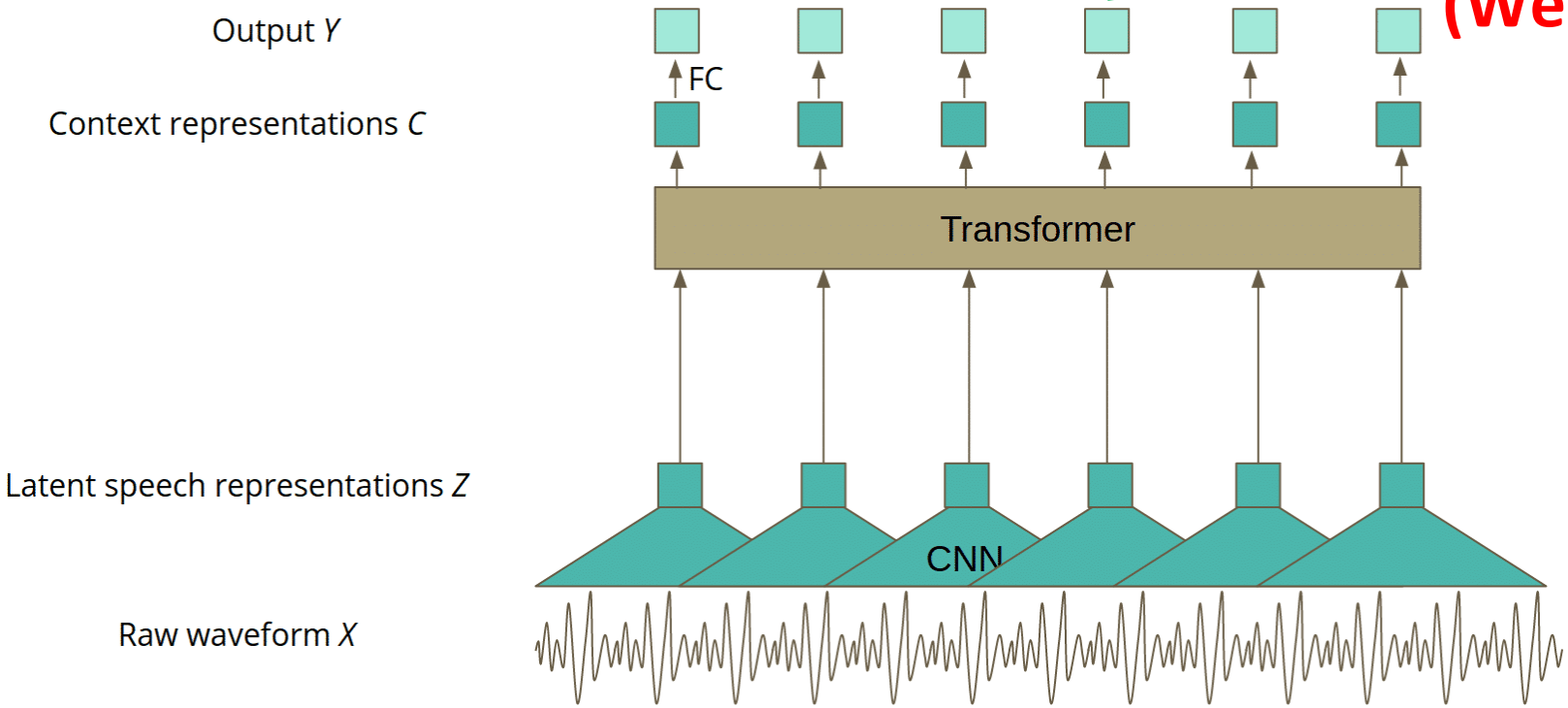
- **Total Loss**

During pre-training, we learn representations of speech audio by solving a contrastive task $\mathcal{L}_m$ which requires to identify the true quantized latent speech representation for a masked time step within a set of distractors. This is augmented by a codebook diversity loss $\mathcal{L}_d$ to encourage the model to use the codebook entries equally often.

$$\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_d$$

Model

Diagnosis of Disease
(We can immediately do it)

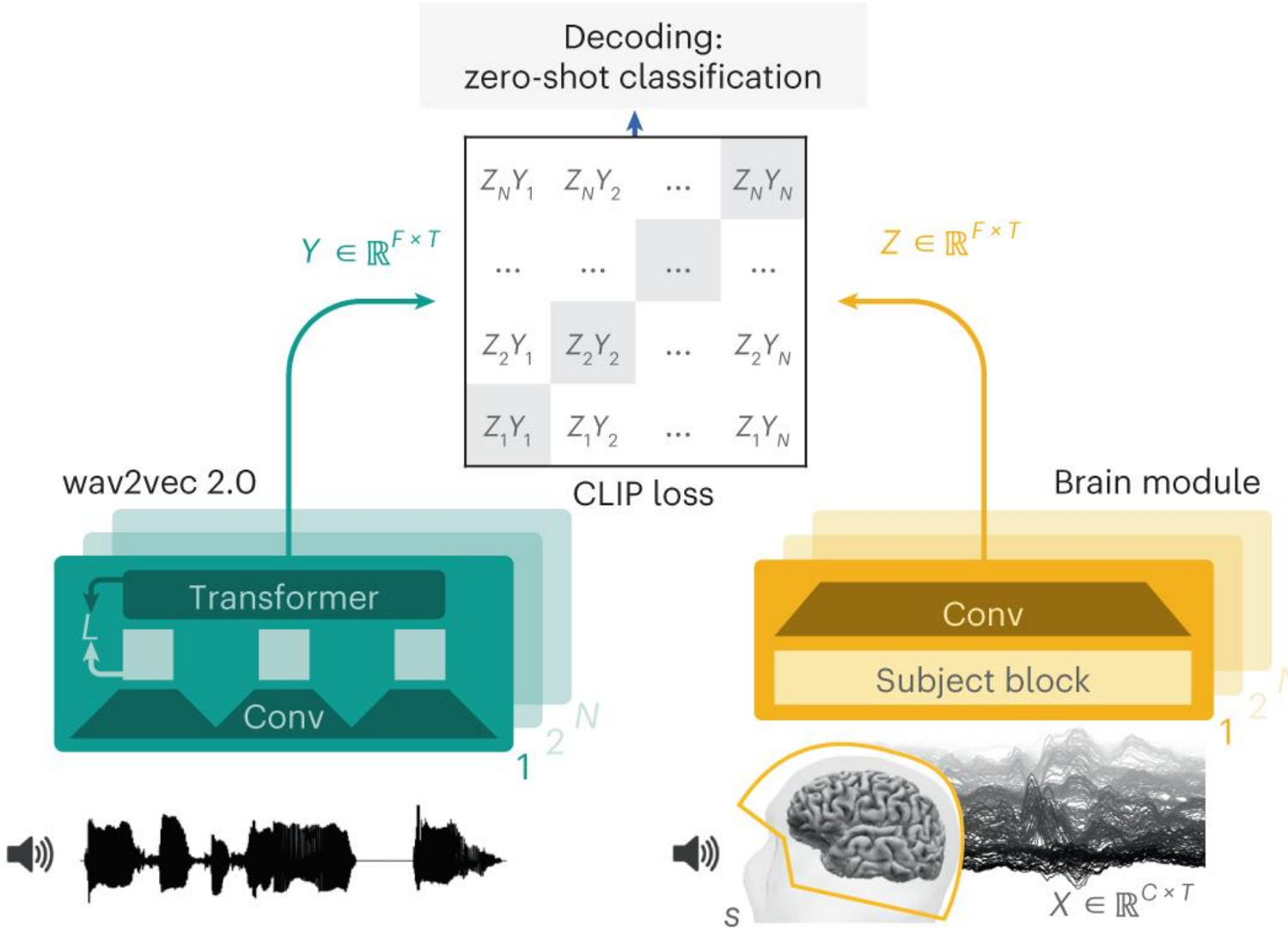


$$g(Z) \rightarrow C(c_1, \ldots, c_T)$$

$$f(\chi) \rightarrow Z(z_1, \ldots, z_T)$$

Decoding speech perception

- Every year, traumatic brain injuries, strokes and neurodegenerative diseases cause thousands of patients lose their ability to speak or even communicate. Brain–computer interfaces (BCIs) have raised high expectations for the detection and restoration of communication abilities in such patients.
- In sum, decoding language from brain activity is, so far, limited either to invasive recordings or to impractical tasks. Interestingly, both of these approaches tend to follow a similar method: that is, (1) training a model on a single participant and (2) aiming to decode a limited set of interpretable features (Mel spectrogram, letters, phonemes, small set of words).
- Instead, here we propose to decode speech from non-invasive brain recordings by using (1) a single architecture trained across a large cohort of participants and (2) deep representations of speech learned with self-supervised learning on a large quantity of speech data



Model approach. We aim to decode speech from the brain activity of healthy participants recorded with MEG or EEG while they listen to stories and/ or sentences. For this, our model extracts the deep contextual representations of 3 s speech signals (Y of F feature by T time samples) from a pretrained speech module' (wav2vec 2.0: ref. 29).

learns the representations (Z) of the brain activity on the corresponding 3 s window (X of C recording channels by T time samples) that maximally align with these speech representations with a contrastive loss.

The representation Z is given by a deep convolutional network. At evaluation, we inpu t the model with left-out sentences and compute the probability of each 3 s speech segment given each brain representation.

The resulting decoding can thus be 'zero shot' in that the audio snippets predicted by the model need not be present in the training set. This approach is thus more general than standard classification approaches where the decoder can only predict the categories learnt during training.

Results

- **Accurately decoding speech from MEG and EEG recordings**

Datasets

Dataset	Language	Type	Sensors	Participants	Duration	Segments	Vocabulary	Segments	Vocabulary	Word overlap (%)
Broderick 2019	English	EEG	128	19	19.2 h	2,645	1,418	1,842	764	67
Brennan and Hale 2019	English	EEG	60	33	6.7 h	1,211	513	190	148	60
Schoffelen 2019	Dutch	MEG	273	96	80.9 h	5,497	1,754	1,270	745	85
Gwilliams 2022	English	MEG	208	27	56.2 h	4,417	1,810	1,363	846	64

Table 2

Model	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
Random model	5.3 ± 0.1	0.5 ± 0.1	0.7 ± 0.1	0.8 ± 0.1
Base model	6.0 ± 0.9	1.0 ± 0.3	12.4 ± 1.2	20.6 ± 1.8
+ Contrastive	8.0 ± 4.8	9.7 ± 1.0	55.1 ± 0.7	55.1 ± 0.9
+ Deep Mel	24.7 ± 3.2	15.4 ± 1.6	64.4 ± 0.8	61.2 ± 0.6
+ wav2vec 2.0	**25.7** ± 2.9	**17.7** ± 0.6	**70.7** ± 0.1	**67.5** ± 0.4

Top-10 segment-level accuracy (%) for a random baseline model that predicts a uniform distribution over the segments ('random'), a convolutional network trained to predict the Mel spectrograms with a regression loss ('base'), the same model trained with a contrastive CLIP loss ('+ Contrastive') and our model, which is trained to predict the features of wav2vec2.0 with a contrastive loss ('+ wav2vec 2.0'). We also report the performance obtained with training, from scratch, a deep learning based speech representation using a contrastive loss ('+ Deep Mel'). Values are mean ±s.d. across three random initializations of the model's weights. The best accuracy across methods is indicated in bold.

From: Decoding speech perception from non-invasive brain recordings

Model	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
Random model	0.5±0.0	0.1±0.0	0.1±0.0	0.1±0.0
Base Model	0.7±0.2	0.1±0.0	3.0±0.3	5.8±0.6
+ Contrastive	0.9±0.6	2.1±0.4	26.2±0.6	25.1±0.6
+ Deep Mel	**5.2**±1.1	4.2±0.7	34.8±1.2	31.6±1.0
+ wav2vec 2.0	**5.2**±0.8	**5.0**±0.4	**41.3**±0.1	**36.8**±0.4

Top-1 Accuracy. Segment-level top-1 accuracy related to Table 2.

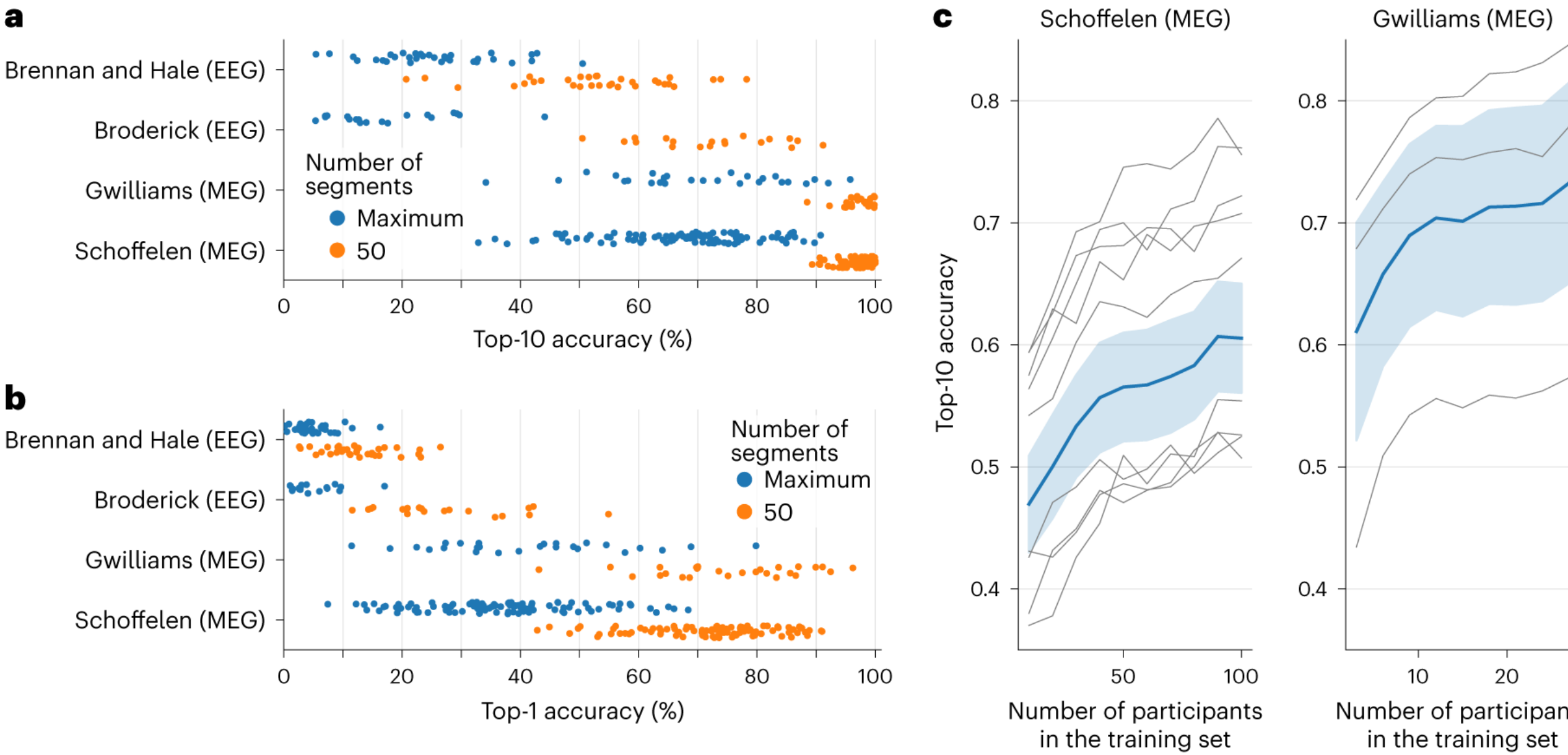

a
Brennan and Hale (EEG)
Broderick (EEG)
Gwilliams (MEG)
Schoffelen (MEG)
Number of segments
Maximum
50
0
20
40
60
80
100
Top-10 accuracy (%)
b
Brennan and Hale (EEG)
Broderick (EEG)
Gwilliams (MEG)
Schoffelen (MEG)
Number of segments
Maximum
50
0
20
40
60
80
100
Top-1 accuracy (%)
c
Schoffelen (MEG)
Gwilliams (MEG)
Top-10 accuracy
0.8
0.7
0.6
0.5
0.4
50
100
10
20
Number of participants in the training set
Number of participan in the training set

Figure 2

Brain module' evaluation

Model	Broderick (EEG)	Brennan (EEG)	Schoffelen (MEG)	Gwilliams (MEG)	delta	p-val
Our model	**17.7** ± 0.6	25.7 ± 2.9	**67.5** ± 0.4	**70.7** ± 0.1		
- Spatial attention dropout	16.0 ± 1.7 *	**26.8** ± 0.7	67.5 ± 0.2	69.0 ± 0.2 *	0.4	0.009
- GELU + ReLU	16.4 ± 0.1	24.6 ± 2.1	65.8 ± 0.6 *	68.8 ± 1.3 *	1.6	$< 10^{-18}$
- Final convs	14.2 ± 1.1 *	19.0 ± 4.4 *	67.5 ± 0.3	68.9 ± 0.9 *	1.1	$< 10^{-10}$
- Non-residual GLU conv	8.4 ± 6.8 *	6.0 ± 0.2 *	67.0 ± 0.2	70.2 ± 0.2	1.6	$< 10^{-10}$
- Skip connections	13.9 ± 2.0 *	24.2 ± 2.7	65.4 ± 0.4 *	66.2 ± 0.3 *	2.4	$< 10^{-21}$
- Initial 1x1 conv	15.4 ± 0.6	22.1 ± 1.9 *	62.9 ± 0.9 *	67.7 ± 0.7 *	3.4	$< 10^{-26}$
- Spatial attention	15.4 ± 0.6 *	20.6 ± 2.2 *	65.9 ± 0.3 *	65.5 ± 0.4 *	2.5	$< 10^{-22}$
- Subj layer	8.1 ± 1.9 *	20.2 ± 1.3 *	42.4 ± 0.1 *	47.0 ± 1.3 *	14.4	$< 10^{-28}$
- Clamping	0.5 ± 0.0 *	14.1 ± 1.0 *	1.5 ± 0.3 *	23.6 ± 24.6 *	26.2	$< 10^{-29}$

To evaluate the elements of the brain module, we performed a series of ablation experiments, and trained the corresponding models on the same data.

Overall, these ablations show that several elements impact performance: performance systematically **decreases when removing skip connections, the spatial attention module, and the initial or final convolutional layers of the brain module**. These results also show the importance of clamping the MEG and EEG signals. Finally, additional experiments show that the present end-to-end architecture is robust to MEG and EEG artefacts, and **requires little preprocessing of the MEG and EEG signals**

Impact of the number of participants

To test whether our model effectively leverages the inter-individual variability, we trained it on a variable number of participants and computed its accuracy on the first 10% of participants. As shown in Fig. 2c, **decoding performance steadily increases as the model is trained with more participants on the two MEG datasets.**

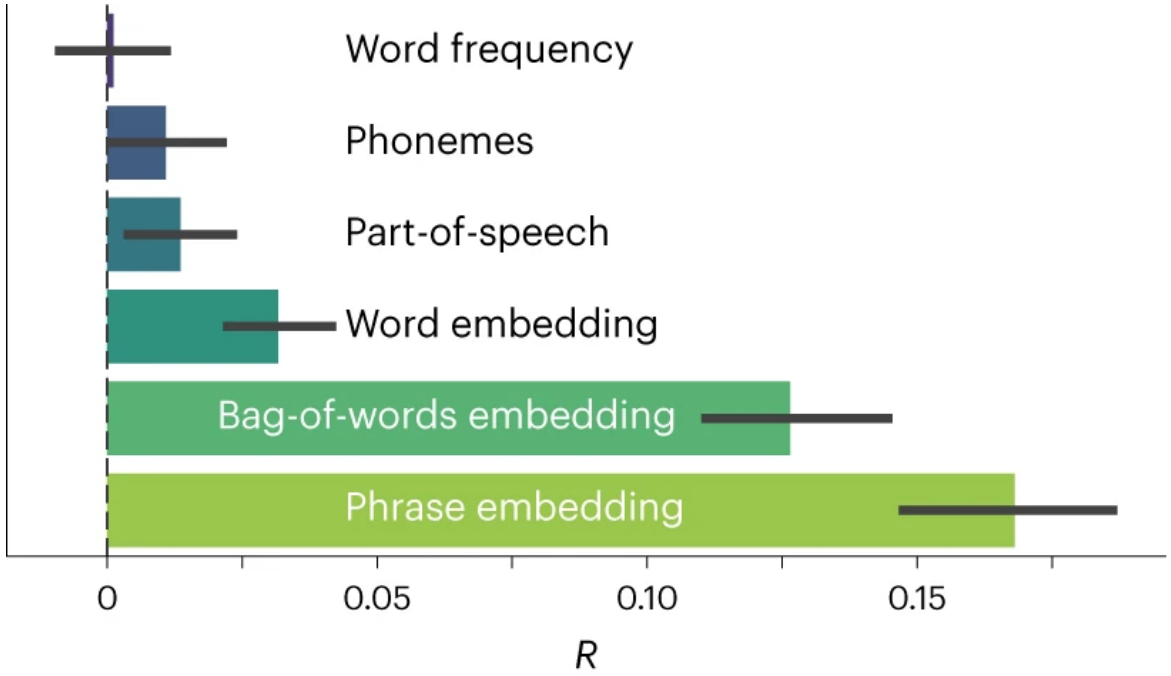
This result shows that our model effectively learns neural representations that are common across participants, while also accommodating participant-specific representations through the participant layer described in Methods

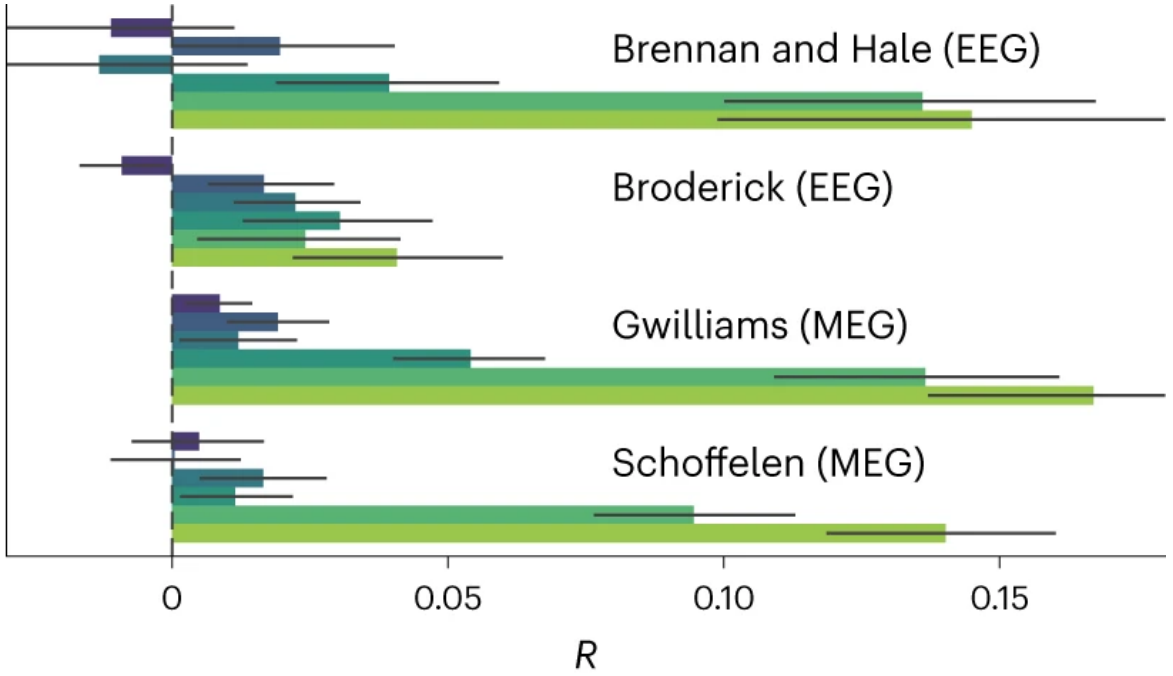
Decoded representations best correlate with phrase embeddings

- What type of representation does our model use to decode speech from brain signals? This interpretability question is notoriously difficult to address


Segment 1
he
the
he
wished
as
wished
rounding
was
he
he
across
were
hurt
of
meeting
were
there
pulled
see
the
there
Segment 2
roots
the
politeness
violating
sucked
the
the
records
stiff
sweet
note
his
strike
sweet
water
hills
water
chord
pointed
of
of
roadside
Segment 3
matter
matter
caught
and
the
him
he
archeology
first
tell
built
one
tell
a
rapped
of
the
journal
you
like
authoritative
vaguely
you
author
because
gray
to
replied
his
nor
the
he
because
there
the
come
Segment 4
just
just
tucker
growing
then
his
then
places
at
a
would
back
his
best
were
horrible
prepare
his
a
of
abandoning
took
pain
wind
tray
says
design
must
Segment 5
air
which
bridge
a
and
sweeping
was
then
air
smelled
smelled
a
planet
bad
himself
that
the
tiny
as
put
in
in
design
way
way

Decoded representations best correlate with phrase embeddings





The R values quantify the extent to which phonemes, word frequency, part-of-speech, word embedding and phrase embedding predict the probability of the predicted word to be correct. Error bars are the s.e.m. across participants

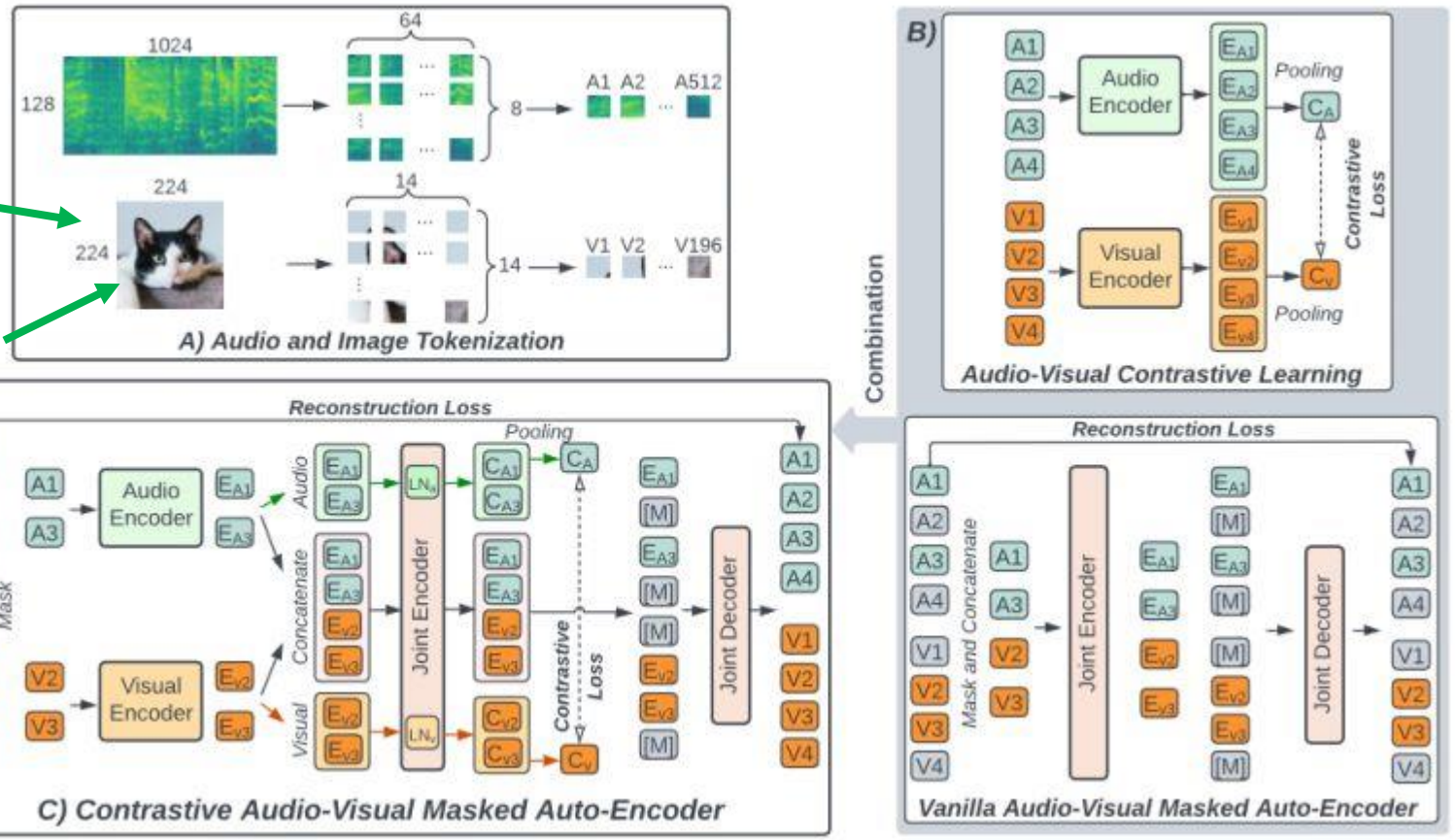
CONTRASTIVE AUDIO-VISUAL MASKED AUTOENCODER

Yuan Gong et al (MIT),April 2023

Code and pretrained models are at https://github.com/yuangongnd/cav-mae.

MEG, EEG or fMRI

Foundation model on Omics



Appendix
Methods

- **Problem formalization**

Aim to decode speech from a time series of high-dimensional brain signals recorded with non-invasive MEG or EEG while healthy volunteers passively listened to spoken sentences in their native language

Let $\boldsymbol{X \in R^{C \times T}}$ be a segment of a brain recording of a given participant while she listens to a speech segment of the same duration, with C the number of MEG or EEG sensors and T the number of time steps.

Let $Y \in R^{F \times T}$ be the latent representation of speech, using the same sample rate as *X* for simplicity, here the Mel spectrogram with *F* frequency bands.

supervised decoding consists of finding a decoding function:

$\boldsymbol{f_{reg}}$: $R^{C \times T} \rightarrow R^{F \times T}$ such that **f**reg predicts *Y* given *X*.

$$\hat{Y} = f_{reg}(X)$$

Is the representation of speech decoded from the brain.

- **When freg belongs to a parameterized family of models like deep neural**

  networks, it can be trained with a regression loss $L_{reg}(Y, \hat{Y})$ (for example, the mean square error)

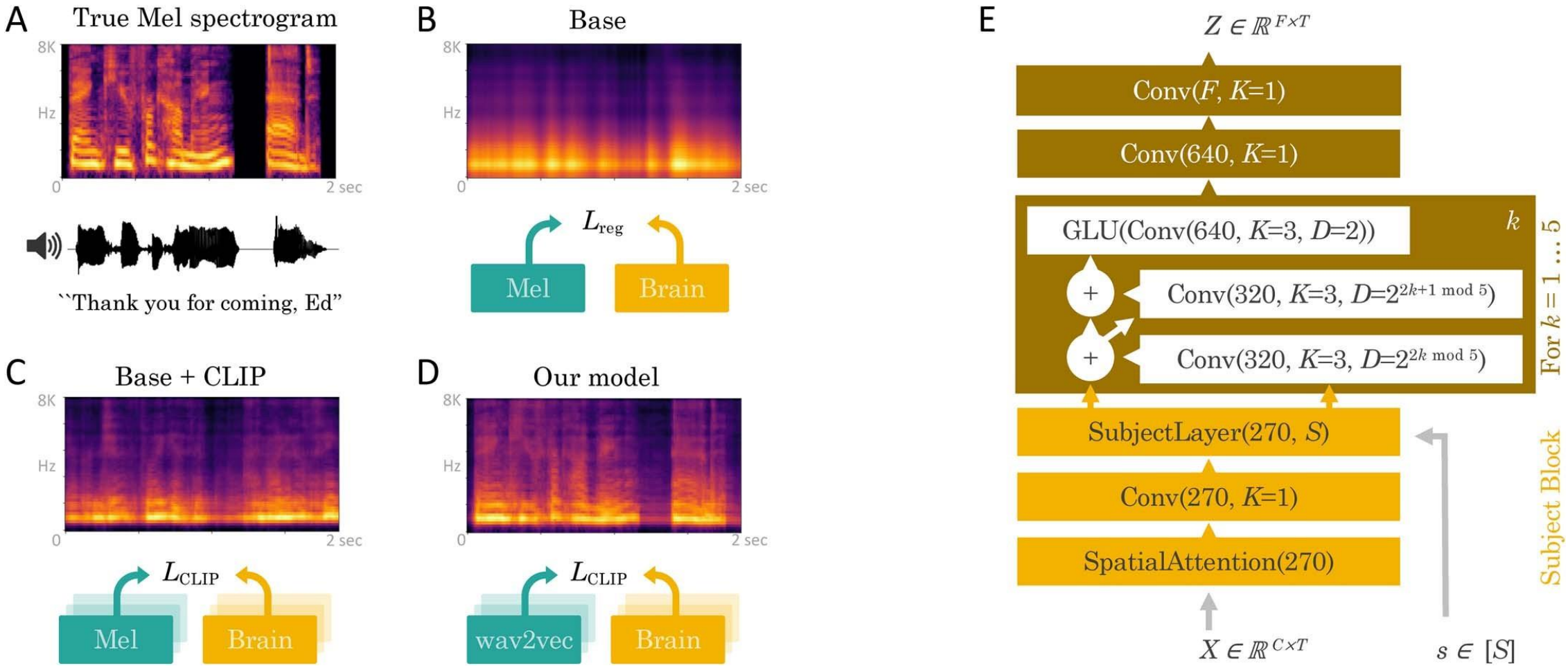
$$\min_{f_{req}} \sum_{X,Y} L_{reg}(Y, f_{reg}(X))$$

This direct regression approach appears to be dominated by a non-distinguishable broadband component when speech is present (Extended Data Fig. 4a,b). This challenge motivates our three main contributions:

the introduction of a contrastive loss,

a pretrained deep speech representation

and a dedicated brain decoder


A
True Mel spectrogram
8K
Hz
0
2 sec
``Thank you for coming, Ed"
B
Base
L_{reg}
Mel
Brain
C
Base + CLIP
L_{CLIP}
Mel
Brain
D
Our model
L_{CLIP}
wav2vec
Brain
E
$Z \in \mathbb{R}^{F \times T}$
Conv(F, K=1)
Conv(640, K=1)
GLU(Conv(640, K=3, D=2))
k
Conv(320, K=3, D=$2^{2k+1 \bmod 5}$)
Conv(320, K=3, D=$2^{2k \bmod 5}$)
For $k = 1 \dots 5$
SubjectLayer(270, S)
Conv(270, K=1)
SpatialAttention(270)
Subject Block
$X \in \mathbb{R}^{C \times T}$
$s \in [S]$

Model

- **Contrastive loss**

We opted for a contrastive objective and thus replaced the regression loss with the 'CLIP' loss (originally for Contrastive Language-Image Pre-Training), which was originally designed to match latent representations in two modalities, text and images.
Unlike the regression objective, this contrastive loss leads the model to find a combination of features that maximally discriminates samples in the batch.

Let *X* be a brain recording segment and $Y \in R^{F \times T}$ the latent representation of its corresponding sound (also known as 'positive sample').
We sample $N-1$ negative samples $\bar{Y}_{j \in \{1,2,\ldots,N-1\}}$ over our dataset and we add the positive sample as $\bar{Y}_N = Y$.
We want our model to predict the probabilities $\forall j \in \{1,\ldots,N\}$, $P_j = P(\bar{Y}_j = Y)$. We thus train a model f_{clip} mapping the brain activity *X* to a latent representation $Z = f_{clip}(X) \in R^{F \times T}$. The estimated probability can then be approximated by the dot product of *Z* and the candidate speech latent representations Y_j, followed by a softmax:

$$\hat{P}_j = \frac{e^{\langle Z, \bar{Y}_j \rangle}}{\sum_{i=1}^{N} e^{\langle Z, \bar{Y}_i \rangle}}$$

with $\langle \cdot, \cdot \rangle$ the inner product over both dimensions of Z and $\bar{Y}$. We then train **f**clip with a cross-entropy between p_j and $\bar{p}_j$. Note that for a large enough dataset, we can neglect the probability of sampling twice the same segment, so that we have $p_j = I_{(j=N)}$, and the cross-entropy simplifies to

$$L_{CLIP}(p, \hat{p}) = -\log \hat{P}_N = -\langle Z, Y \rangle + \log \left(\sum_{i=1}^{N} e^{\langle Z, \bar{Y}_i \rangle} \right)$$

- **Brain module**

For the brain module, we introduce a deep neural network f_{clip}, input with raw MEG and EEG times series *X* and a one-hot encoding of the corresponding participant *s*, and outputs the latent brain representation *Z*, with the same sample rate as *X*. This architecture consists of (1) a spatial attention layer over the MEG and EEG sensors followed (2) by a participant-specific 1 × 1 convolution designed to leverage inter-individual variability, which input to (3) a stack of convolutional blocks. An overview of the model is given in the Extended Data Fig. 4e. In the following, given a tensor *U*, we note *U*(*i*,…) access to specific entries in the tensor.

- **Spatial attention.**

The brain data are first remapped onto $D1 = 270$ channels with a spatial attention layer based on the location of the sensors. The three-dimensional sensor locations are first projected on a two-dimensional plane obtained with the MNE-Python function find_layout 45, which uses a device-dependent surface designed to preserve the channel distances. Their two-dimensional positions are finally normalized to [0, 1]. For each output channel, a function over $[0,1]^2$ is learnt, parameterized in the Fourier space. The weights over the input sensors are then given by the softmax of the function evaluated at the sensor locations. Formally, each input channel i has a location (x_i, y_i) and each output channel j is attached a function a_j over $[0,1]^2$, parameterized in the Fourier space as $z_j \in C^{K\times K}$ with $K = 32$ harmonics along each axis, that is

$$a_j(x,y) = \sum_{k=1}^{K}\sum_{l=1}^{K}\left[Re\left(z_j^{(k,l)}\right)cos(2\pi(kx+ly)) + Im\left(z_j^{(k,l)}\right)sin(2\pi(kx+ly))\right]$$

The output is given by a softmax attention based on the evaluation of a_j at each input position (x_i, y_i):

$$\forall j \in \{1,\dots,D_1\} SA^{(j)} \frac{1}{\sum_{i=1}^{C} e^{a_j(x_i,y_i)}} \left(\sum_{i=1}^{C} e^{a_j(x_i,y_i)} X^{(i)}\right)$$

with SA the spatial attention. In practice, as a_j is periodic, we scale down (*x*, *y*) to keep a margin of 0.1 on each side. We then apply a spatial dropout by sampling a location $\left(x_{drop}, y_{drop}\right)$ and removing from the softmax each sensor that is within a distance of $d_{drop} = 0.2$ of the sampled location.

The initial motivation for spatial attention was to allow for a cross-dataset model to be defined in a way that would generalize across a diverse number location and set of sensors. Interestingly, we observed this layer to introduce an inductive bias that is beneficial to the prediction accuracy (Extended Data Fig. 2). See Extended Data Fig. 4 for a visualization of the learnt attention maps over each dataset. We then add a 1 × 1 convolution (that is, with a kernel size of 1) without activation and with the same number *D*1 of output channels.

- **Participant layer.**

  To leverage inter-individual variability, we learn a matrix $M_s \in \mathbb{R}D1, D1$ for each participant $s \in [S]$ and apply it after the spatial attention layer along the channel dimension. This is similar to but more expressive than the participant embedding used by ref. 46 for MEG encoding, and follows decade of research on participant alignment

- **Residual dilated convolutions.**

  We then apply a stack of five blocks of three convolutional layers. For the *k*th block, the first two convolutions are applied with residual skip connections (except for the very first one where the number of dimension potentially doesn't match), outputs $D2 = 320$ channels and are followed by batch normalization and a GELU (Gaussian Error Linear Unit) activation.

  The two convolutions are also dilated to increase their receptive field, by $2^{k mod\ 5}$ and $2^{(2k+1) mod\ 5}$ (with k zero indexed), respectively. The third layer in a block outputs $2D_2$ channels and uses a GLU (Gated Linear Unit) activation, which halves the number of channels. All convolutions use a kernel size of 3 over the time axis, a stride of 1 and sufficient padding to keep the number of time steps constant across layers.

The output of the model is obtained by applying two final 1 × 1 convolutions: first with $2D2$ outputs, followed by a GELU and finally with F channels as output, thus matching the dimensionality of speech representations. Given the expected delay between a stimulus and its corresponding brain responses, we further shift the input brain signal by 150 ms into the future to facilitate the alignment between Y and Z. The impact of this offset is considered in the Supplementary Section A.5.

- **Speech module**

  The Mel spectrogram is a low-level representation of speech inspired from the cochlea and is thus unlikely to match the rich variety of cortical representations. Consequently, we replaced the Mel spectrograms with latent representations of speech. For this, we propose either to learn these representations end-to-end (‘Deep Mel’ model) or to rely on those learnt by an independent self-supervised speech model (wav2vec 2.0; ref. 29).

- **End-to-end speech representations with Deep Mel**

The ‘Deep Mel’ module uses the same deep convolutional architecture to the brain module devoid of the participant block, and thus simultaneously learns to extract speech and MEG and EEG representations such that they are maximally aligned. By definition, and unlike wav2vec 2.0, Deep Mel sees only the audio used in the MEG and EEG datasets. As this end-to-end approach proved to be less efficient than its pretrained counterpart based on wav2vec 2.0, we will thereafter focus on the latter.

- **Pretrained speech representations with wav2vec 2.0.**

  Wav2vec 2.0 is trained with audio data only to transform the raw waveform with convolutional and transformer blocks to predict masked parts of its own latent representations. A previous study showed that the resulting model can be efficiently fine-tuned to achieve state-of-the-art performance in speech recognition. Besides, this model effectively encodes a wide variety of linguistic features. In particular, recent studies have shown that the activations of wav2vec 2.0 linearly map onto those of the brain. Consequently, we here test whether this model effectively helps the present decoding task. In practice, we use the wav2vec2-large-xlsr-53 (ref. 56), which has been pretrained on 56,000 hours of speech from 53 different languages.

Datasets

We test our approach on four public datasets, two based on MEG recordings and two based on EEG recordings. All datasets and their corresponding studies were approved by the relevant ethics committee and are publicly available for fundamental research purposes. We provide an overview of the main characteristics of the datasets in Table 1, including the number of training and test segments and vocabulary sizes over both splits.

Dataset	Language	Type	Sensors	Participants	Duration	Segments	Vocabulary	Segments	Vocabulary	Word overlap (%)
Broderick 2019	English	EEG	128	19	19.2 h	2,645	1,418	1,842	764	67
Brennan and Hale 2019	English	EEG	60	33	6.7 h	1,211	513	190	148	60
Schoffelen 2019	Dutch	MEG	273	96	80.9 h	5,497	1,754	1,270	745	85
Gwilliams 2022	English	MEG	208	27	56.2 h	4,417	1,810	1,363	846	64

For all datasets, healthy adult volunteers passively listened to speech sounds (accompanied by some memory or comprehension questions to ensure participants were attentive), while their brain activity was recorded with MEG or EEG. In Schoffelen et al.[32], Dutch-speaking participants listened to decontextualized Dutch sentences and word lists (Dutch sentences for which the words are randomly shuffled). The study was approved by the local ethics committee (the local Committee on Research Involving Human Subjects in the Arnhem–Nijmegen region). The data are publicly and freely available after registration on the Donders Repository. In Gwilliams et al. , English-speaking participants listened to four fictional stories from the Masc corpus in two identical sessions of 1 hour. The study was approved by the institutional review board ethics committee of New York University Abu Dhabi. In Broderick et al. , English-speaking participants listened to extracts of *The Old Man and the Sea*. The study was approved by the ethics committees of the School of Psychology at Trinity College Dublin and the Health Sciences Faculty at Trinity College Dublin. In Brennan and Hale, English-speaking participants listened to a chapter of *Alice in Wonderland*. See Supplementary Section A.1 for more details. The study was approved by the University of Michigan Health Sciences and Behavioral Sciences institutional review board (HUM00081060).