Toward a realistic model of speech processing in the brain with self-supervised learning

Juliette Millet⋆ [1,2,3] **Charlotte Caucheteux**⋆ [1,4] **Pierre Orhan**[2] **Yves Boubenec**[2]
Alexandre Gramfort[4] **Ewan Dunbar**[2,5] **Christophe Pallier**[6] **Jean-Rémi King**[1,2]

[1]Meta AI, Paris, France [2]Ecole Normale Supérieure, PSL University, Paris, France
[3]LPI, Université de Paris cité, Paris, France
[4]Université Paris-Saclay, Inria, CEA, Palaiseau, France
[5]University of Toronto, Toronto, Canada
[6]Cognitive Neuroimaging Unit, INSERM, Gif-sur-Yvette, France

juliette.millet@cri-paris.org
ccaucheteux@fb.com
jeanremi@fb.com

Abstract

Several deep neural networks have recently been shown to generate activations similar to those of the brain in response to the same input. These algorithms, however, remain largely implausible: they require (1) extraordinarily large amounts of data, (2) unobtainable supervised labels, (3) textual rather than raw sensory input, and / or (4) implausibly large memory (e.g. thousands of contextual words). These elements highlight the need to identify algorithms that, under these limitations, would suffice to account for both behavioral and brain responses. Focusing on the issue of speech processing, we here hypothesize that self-supervised algorithms trained on the raw waveform constitute a promising candidate. Specifically, we compare a recent self-supervised architecture, Wav2Vec 2.0, to the brain activity of 412 English, French, and Mandarin individuals recorded with functional Magnetic Resonance Imaging (fMRI), while they listened to ≈ 1 h of audio books. Our results are four-fold. First, we show that this algorithm learns brain-like representations with as little as 600 hours of unlabelled speech – a quantity comparable to what infants can be exposed to during language acquisition. Second, its functional hierarchy aligns with the cortical hierarchy of speech processing. Third, different training regimes reveal a functional specialization akin to the cortex: Wav2Vec 2.0 learns sound-generic, speech-specific and language-specific representations similar to those of the prefrontal and temporal cortices. Fourth, we confirm the similarity of this specialization with the behavior of 386 additional participants. These elements, resulting from the largest neuroimaging benchmark to date, show how self-supervised learning can account for a rich organization of speech processing in the brain, and thus delineate a path to identify the laws of language acquisition which shape the human brain.

1 Introduction

The performance of deep neural networks has taken off over the past decade. Algorithms trained on object classification, text translation, and speech recognition are starting to reach human-level performance Xu et al. [2021]. Furthermore, the *representations* of these algorithms have repeatedly shown to correlate with those of the brain Kriegeskorte [2015], Yamins and DiCarlo [2016], Kietzmann et al.

Preprint. Under review.

[2018], Kell and McDermott [2019], Cichy and Kaiser [2019], Toneva and Wehbe [2019], Millet and King [2021], Caucheteux and King [2022], suggesting that these algorithms converge to brain-like computations.

Such convergence, however, should not obscure the major differences that remain between these deep learning models and the brain. In particular, the above comparisons derive from models trained with (1) extraordinarily large amounts of data (40GB for GPT-2 Radford et al. [2019], the equivalent of multiple lifetimes of reading), (2) supervised labels which are rare in human experience (e.g. Yamins and DiCarlo [2016]), (3) data in a textual rather than a raw sensory format, and/or (4) considerable memory (e.g., language models typically have parallel access to thousands of context words to process text). These differences highlight the pressing necessity to identify architectures and learning objectives which, subject to these four constraints, would be sufficient to account for both behavior and brain responses.

Here, we hypothesize that the latest self-supervised architectures trained on raw sensory data constitute promising candidates Borgholt et al. [2022], Bardes et al. [2021], Baevski et al. [2020]. We focus on Wav2Vec 2.0 Baevski et al. [2020], an architecture that stacks convolutional and transformer layers to predict a quantization of the latent representations of speech waveforms. We train Wav2Vec 2.0 on 600 h of speech—a quantity roughly comparable to what infants are exposed to during early language acquisition Dupoux [2018].

We use standard encoding analyses Naselaris et al. [2011], Huth et al. [2016], Yamins and DiCarlo [2016], Kell et al. [2018] (see Figure 1) to compare this model to the brains of 412 healthy volunteers (351 English speakers, 28 French speakers, and 33 Mandarin speakers) recorded with functional magnetic resonance imaging (fMRI) while they passively listened to approximately one hour of audio books in their native language Nastase et al. [2020], Li et al. [2021] (8.5 hours of distinct audio materials in total).

To better understand the similarities between Wav2Vec 2.0 and the brain, we compare brain activity to each layer of this model, as well as to several variants, namely (1) a random (untrained) Wav2Vec 2.0 model, (2) a model trained on 600 h of non-speech sounds, (3) a model trained on 600 h of non-native speech (for example, a model trained on English speech and mapped onto the brain responses to French-speaking participants), (4) a model trained on 600 h of native speech (for example, a model trained on English speech and mapped onto the brain responses to English participants), and (5) a model trained directly on speech-to-text (i.e., a supervised learning scheme) on the native language of the participants.

Our results provide four main contributions. First, self-supervised learning leads Wav2Vec 2.0 to learn latent representations of the speech waveform similar to those of the human brain. Second, the functional hierarchy of its transformer layers aligns with the cortical hierarchy of speech in the brain, and reveals the whole-brain organisation of speech processing with an unprecedented clarity. Third, the auditory-, speech-, and language-specific representations learned by the model converge to those of the human brain. Fourth, behavioral comparisons to 386 supplementary participants' results on a speech sound discrimination task confirm this common language specialization.

2 Methods

2.1 Models

We train several variants of Wav2Vec 2.0 Baevski et al. [2020] from scratch on different datasets using two different learning objectives (a self-supervised and a supervised objective).

2.1.1 Architecture

Wav2Vec 2.0 consists of three main modules. First, a feature encoder composed of seven blocks of temporal convolutions (output dimension 512) transforms the speech input S (raw mono waveform at 16 kHz) into a latent representation z (output dimension of 512, frequency 49 Hz, stride of 20 ms between each frame, receptive field of 25 ms). Second, a quantization module discretizes z into q, a dictionary of discrete and latent representations of sounds. Third, z is input to a "context network" consisting of 12 transformer blocks (model dimension 768, inner dimension 3072, and 8 attention heads), which together yield a contextualized embedding c, of the same dimensionality of q.

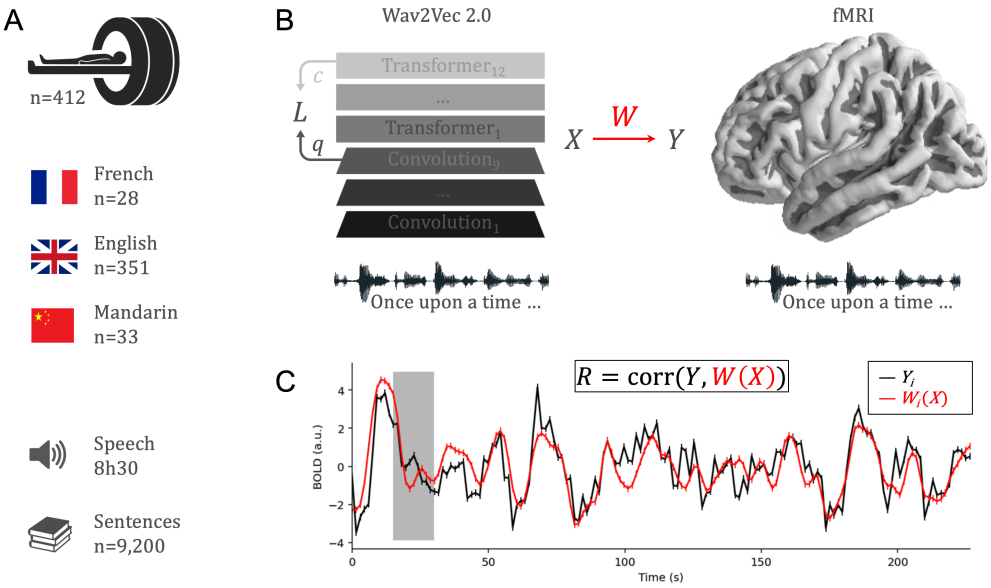

Figure 1: **Comparing speech representations in brains and deep neural networks. A.** We analyse the brain activity of 412 participants recorded with functional Magnetic Resonance Imaging (fMRI) while they passively listened to audio books in their native language (French, English or Mandarin). **B.** After training Wav2Vec 2.0 Baevski et al. [2020] with self-supervised learning (L) over 600 h of unlabelled speech, we extract its activations in response to the audio books that were presented to the participants. We assess the similarity between the activations of the model X and brain activity Y with a standard encoding model W Nastase et al. [2020] evaluated with a cross-validated Pearson correlation R. **C.** Example of the true BOLD response (black) and the predicted BOLD response (red) estimated from a linear projection of the model's activations in a single voxel of the auditory cortex for the first 200 s of a representative story in the test set.

2.1.2 Learning objective

Self-supervised learning. In this training paradigm, the model optimizes two losses. The first loss is contrastive and requires the model to predict the quantized representation q of some masked input using c, from a finite set of quantized representations drawn from the input sample. The second loss ensures that the quantized representations are diverse. See A.2 and Baevski et al. [2020] for details.

Supervised learning. In this training paradigm, the quantization module is discarded and a linear layer mapping c to phonemes is added at the end of the pipeline. The model is randomly initialized and all layers (including the feature encoder) are trained using a Connectionist Temporal Classification (CTC) Graves [2012] loss to perform phone recognition.

For both training paradigms, we extract the activations of each layer from both the feature encoder (outputting z) and the context network (outputting c). Thus, we study both the convolutional and transformer blocks. We extract these internal representations using input windows consisting of 10 s of raw waveform, with a stride of 5 s.

2.1.3 Training

Datasets. We successively train different Wav2Vec 2.0 models using each of four datasets: (i) the French and (ii) English CommonVoice corpora Ardila et al. [2020], (iii) the MAGICDATA Mandarin Chinese Read Speech Corpus Co. [2019], and (iv) a non-speech subset of the Audioset dataset Gemmeke et al. [2017], which contains recordings of various acoustic scenes.

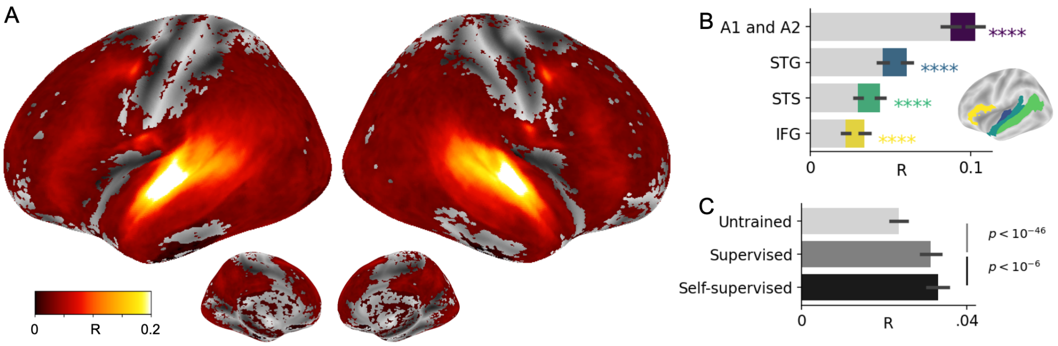

Figure 2: **Self-supervised learning suffices for Wav2Vec 2.0 to generate brain-like representations of speech. A.** Neural predictivity score (R) assessed for each subject and voxel independently, and here averaged across subjects for clarity. Only scores significantly above chance level, as assessed using a two-sided Wilcoxon test across subjects after correction for multiple comparison are color-coded ($p < 10^{-10}$). **B.** R scores for the same Wav2Vec2 model, averaged across subjects and voxels in four brain areas typically involved during speech processing (the primary and secondary auditory cortices, the superior temporal gyrus, the superior temporal sulcus, and the infero-frontal gyrus). In grey, the brain score obtained with a randomly initialized Wav2Vec 2.0 architecture. Error bars are the standard errors of the mean (SEM) across subjects. The stars indicate a significant difference between the random and trained model (all $p < 10^{-4}$). **C.** R scores of Wav2Vec 2.0 without training (top), trained with a supervised (middle) and self-supervised learning rule (bottom), on the same 600 hours of speech. Scores are averaged across subjects and voxels and error bars are SEM across subjects.

Preprocessing. All the audio datasets were randomly subsampled to have an approximate size of 600 hours, downsampled to 16 kHz and converted to mono with the Sox software[1]. We randomly split the datasets into a training (80%), a validation (10%) and a test set (10%). The audio recordings we use from the Audioset dataset are filtered so that they do not contain speech or any sounds produced by humans, such as laughter or singing. For the speech datasets, we also use their corresponding annotations (in the supervised settings). We phonemize these annotations using eSpeakNG[2]. Note that the number of different phoneme symbols in these annotations is similar for French (32), English (39), and Mandarin Chinese (33).

Implementation. We train all of our models using the fairseq implementation of Wav2Vec 2.0[3] using default hyperparameters. We also analyze a model whose parameters were randomly initialised ("untrained" model).

We use self-supervised learning to train four models: three on the speech datasets (French, English, and Mandarin) and one on the acoustic scenes dataset. In each case, the training was performed using the same configuration file (namely, the base configuration provided in the fairseq repository for pretraining Wav2Vec 2.0 on LibriSpeech Panayotov et al. [2015]). We train the models for 400k updates and select the ones with the best validation loss.

We use the supervised training paradigm to train three models, on the French, English, and Mandarin datasets, respectively. Each training was performed using the same configuration file, which was identical to the configuration provided in the fairseq repository for fine-tuning Wav2Vec 2.0 on the 960 hour Voxpopuli corpus Wang et al. [2021], except that parameters were not frozen (`freeze_finetune_updates`= 0) and learning was performed on all parameters of the models using the CTC loss (`feature_grad_mult`= 0.1). We train the models for 400k updates and we use the ones with the best word error rate (WER) on the validation set. The French model obtains 13.9 WER, the English model 28.6 WER, and the Mandarin model 4.6 WER, on their respective test sets.

[1] http://sox.sourceforge.net/

[2] https://github.com/espeak-ng/espeak-ng

[3] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

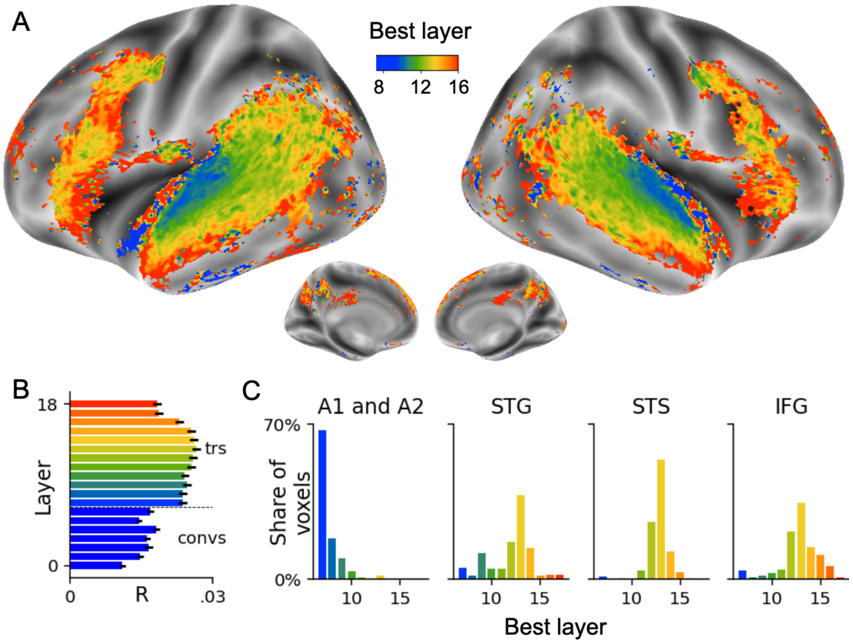

Figure 3: **The functional hierarchy of Wav2Vec 2.0 maps onto the speech hierarchy in the brain.** **A.** We compute the R score for each layer of Wav2Vec 2.0 separately and estimate, for each voxel, the layer with highest brain score on average across subjects. Only the voxels with significant brain scores are displayed ($p < 10^{-18}$). While the first transformer layers (blue) map onto the low-level auditory cortices (A1 and A2), the deeper layers (orange and red) map onto brain regions associated with higher-level processes (e.g. STS and IFG). **B.** Layer-wise R scores averaged across all voxels. Error bars are SEM across subjects. **C.** Proportion of voxels with most predictive layer (x-axis) in four regions typically involved in speech processing. While most voxels in the primary cortex are best predicted by the first layers of the transformer, higher-level brain areas are best predicted by deeper layers.

2.2 Functional MRI

We analyse a composite set of fMRI recordings aggregated from the *Little Prince* Li et al. [2021] and the *Narratives* public datasets Nastase et al. [2020].

Narratives. This is a public dataset consisting of the fMRI recordings of 345 native English-speaking participants listening to English narratives (4.6 hours of unique audio in total). Following the original manuscript and repository Nastase et al. [2020], we (1) focus on fifteen representative stories and ignore the narratives that have been modified by scrambling and (2) exclude eight participants because of noisy recordings. Overall, this selection results in a curated dataset of 303 participants listening to fifteen stories ranging from 3 min to 56 min, for a total of 4 hours of unique audio, 36018 words, and 4004 unique words.

The Little Prince. This dataset contains the fMRI data acquired for 48 English native speakers, 33 Mandarin native speakers, and 28 French native speakers, each listening to *The Little Prince* in their respective native language. The experiment itself was divided into nine runs of approximately 10 min of passive listening. For each language condition, the story was read by a single native speaker. The English, Mandarin, and French audiobooks last 94, 90 and 97 minutes respectively.

Preprocessing. For Narratives, we did not perform additional preprocessing: we use the public preprocessing of the dataset already projected on the surface space ("fsaverage6") without spatial smoothing (labelled "afni-nosmooth" in the data repository). In contrast, the *Little Prince* dataset is only provided in a volumetric space. Consequently, for each language condition separately, we subselected the cortical voxels by computing a brain mask using the average of all participants'

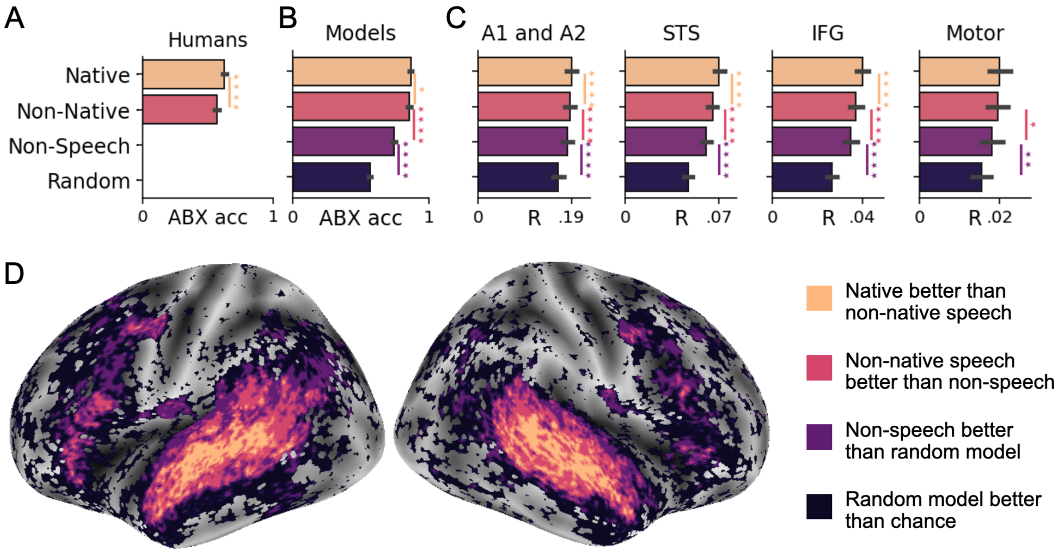

Figure 4: **The specialization of Wav2Vec 2.0's representations follows and clarifies the acoustic, speech, and language regions in the brain.** **A)** We first evaluate humans' language specificity by quantifying their ability to perceive phonemes of their native or non-native languages (Section 2.4) in a ABX matching-to-sample task Schatz [2016] (higher is better). As expected, humans are better at matching phonemes of their native language. **B)** Then, we train four Wav2Vec 2.0 models with self-supervised learning on four datasets – non-speech acoustic scenes, English, and French, and compute their ABX accuracy on the same speech datasets as humans. The 'random' model is Wav2Vec 2.0 without any training. **C)** Neural Predictivity score (R) of each model (with an added model trained on Mandarin), averaged across voxels, in four regions of the brain (Section 2.2). **D)** Acoustic, speech and language specificity for each voxel. For instance, one voxel is considered specific to the 'native' model if its native R score is higher than its 'non-native' R score ($p < .05$). Only the voxels with significant R scores for the untrained model are displayed ($p < 10^{-18}$). Error bars are the SEM across phone pairs in A and B, and across subjects in C. The stars indicates a significant difference between two conditions (Section 2.6).

fMRI data realigned onto a common template brain via Freesurfer Fischl [2012]. These voxels are then projected onto a brain surface using nilearn's `vol_to_surf` function with defaults parameters Abraham et al. [2014] and a 'fsaverage6' template surface Fischl [2012].

For both *Narratives* and *The Little Prince*, fMRI signals are normalized across the time dimension to have a mean of 0 and a variance of 1, for each participant, surface voxel and session independently.

Brain parcellation. For the purposes of certain analyses, we group the fMRI voxels into regions of interest using the Destrieux Atlas Destrieux et al. [2010]. This parcellation results in 75 brain regions in each hemisphere. For simplicity, we label the regions as follows: **A1 and A2** represents Heschl gyrus, which is the anatomical location of the primary and secondary auditory cortices, **STG** and **STS** are the superior temporal gyrus and sulcus, and **IFG** is the inferior frontal gyrus.

2.3 Neural Predictivity score (R)

To quantify the similarity between the network's activations X and the brain recordings Y, we use a standard linear encoding model Huth et al. [2016], Yamins and DiCarlo [2016]. For each subject, we split the data into train and test sets using a five-fold cross-validation setting. For each train split, a linear mapping W is fitted to predict the brain response Y_{train} given X_{train}. W combines a temporal alignment function with fixed weight, and a trained penalized linear regression.

Temporal alignment. The sampling frequency of the model's activations (between 49 and 200 Hz) differs from the sampling frequency of fMRI BOLD signals (0.5 Hz). Furthermore, the BOLD signals have delayed responses spanning over several seconds. Thus, we first convolve the model activations with a standard hemodynamic response function (HRF) using nistats Abraham et al. [2014] `compute_regressor` function with the 'glover' model and default parameters. This results in the convolved activations X'_{train} with the same sampling frequency as the fMRI Y_{train} (see A.3).

Penalised linear regression. Once temporally aligned, we fit an ℓ_2-penalised linear regression that predicts the brain signals Y_{train} given the activations X_{train}. We use the `RidgeCV` function from scikit-learn Pedregosa et al. [2011], with the penalization hyperparameter λ varying between 10 and 10^8 (20 values scaled logarithmically) chosen independently for each dimension with a nested cross-validation over the training set (see A.4).

Evaluation. We evaluate the linear mapping W on the held out sets by measuring Pearson's correlation between predicted and actual brain responses:

$$R = \operatorname{corr}\big(Y_{\text{test}}, W \cdot X_{\text{test}}\big) \quad . \tag{1}$$

Finally, we average the correlation scores across test splits to obtain the final Neural Predictivity score (R).

2.4 Behavioral experiment

To study the representation specific to each language, we compare the way humans and models are able to discriminate native and non-native speech sounds. We use "ABX" tasks, which are discrimination tests during which participants have to listen to three speech sounds: A (e.g., /pop/), B (e.g., /pap/) and X (e.g., /pop/) Schatz [2016]. X is of the same category as A or B, and the participant has to decide which of A or B is the closest to X. Once the results are aggregated across participants of one given language group, the discrimination accuracy indicates whether they can differentiate the categories A and B belong to (in our example, whether they can differentiate the phone pair /o/-/a/).

For this, we use the Perceptimatic human benchmark[4] that contains the ABX results of French and English-speaking participants on speech stimuli from various languages. We focus on the French and English stimuli, which represents ≈ 6,000 ABX triplets (testing 508 English and 524 French phone pairs), with 386 participants in total (193 from each language group).

In Figure 4-A, we report the ABX accuracy of English- and French-speaking participants in both their native and non-native language (either English or French). We first average results per phone pair, and then average over phone pairs to obtain the ABX discrimination accuracy. Similarly, in Figure 4-B, we compute the ABX accuracy of our Wav2Vec 2.0 models on the same evaluation sets as the participants, using the parameters described in Millet and Dunbar [2022]. English and French models are evaluated on the same ('native') or different ('non-native') language stimuli as their training. The untrained and non-speech models are evaluated on both French and English speech stimuli.

2.5 Layer specificity

To report the average layer k^* with the highest brain score for each voxel (Figure 3) while being robust to regression-to-the-mean biases, we first find the best layer k_s for each subject s and each voxel independently and then compute a circular mean across participants: $k^* = \text{angle}\left(\frac{1}{N}\sum_{s=1}^{N}\exp\left(\frac{2i\pi k_s}{K+1}\right)\right)$ with $K = 19$ layers, and $N = 412$ participants.

2.6 Statistics

We assess the reliability of brain scores with second-level analyses across participants, by applying a Wilcoxon signed-rank test across participants. Thus, the resulting p-values are not affected by fMRI auto-correlation within participants. We perform statistical correction for multiple comparisons with Benjamini–Hochberg False Discovery Rate (FDR) across voxels Benjamini [2010].

[4] `https://docs.cognitive-ml.fr/perceptimatic/`

3 Results

Wav2Vec 2.0 maps onto brain responses to speech. We estimate whether the activations of Wav2Vec 2.0 models linearly map onto the human brain activity of 412 individuals listening to audio books in the fMRI scanner. For this, we first independently train three models with 600 h of French, English, or Mandarin, respectively, and compute the Neural Predictivity score (R) with the corresponding participants. Specifically, we (1) convolve the activations (X) of the model with a hemodynamic response function (HRF), (2) train a ℓ_2-penalized linear regression on a training split to map them to brain activity Y, and (3) compute the Pearson correlation coefficient between (i) the true fMRI activity and (ii) the predicted activations on a test split. The models' activations significantly predict brain activity in nearly all cortical areas, reaching the highest R scores in the primary and secondary auditory cortices (Figure 2-A B). These scores are significantly higher than those obtained with a randomly initialised model ($p < 10^{-50}$ on average across voxels), and this comparison is robust across language groups (all $p < 10^{-5}$).

Comparison of self-supervised to supervised models. Does self-supervision reach representations that are as brain-like as those obtained with supervised learning? To address this issue, we trained Wav2Vec 2.0 with an alternative, supervised objective, namely, predicting phonetic annotations from the same 600 hours of speech sounds. We then implemented the R score analyses described above. The results show that self-supervised learning in fact leads to modestly but significantly better R scores than supervised learning (Figure 2-C): $\Delta R = 0.002, p < 10^{6}$.

The hierarchy of Wav2Vec 2.0 maps onto the hierarchy of the cortex. To compare the speech hierarchy in the brain with the functional hierarchy learned by Wav2Vec 2.0, we compare the R score of each layer of the model (Figure 3). First, we observe that convolutional layers are less predictive than transformer layers. Second, within the transformers, the hierarchy of representations aligns with the expected cortical hierarchy Hickok and Poeppel [2007]: while low-level areas (A1, A2) are best predicted by the first transformer layers, higher level areas (IFG, STS) are best predicted by deeper layers. Remarkably, this hierarchy extends to supplementary motor and motor areas in both hemispheres (Figure 3-A).

Language specificity in phone discrimination tasks. The acoustic features underlying speech (fricatives, vowels, and so on) may also characterize non-speech sounds (the sound of tree leaves in the wind, of a stone falling, and so on). Does the model show commonalities merely with general auditory processing in the brain, or does it capture speech-specific processing? If so, does it show commonalities with brain representations that are specific to the native language of the participants, or merely to general speech processing? We first evaluate the specialization of humans' perception to their native language using an ABX behavioral task (Section 2.4). Specifically, we compare 386 French and English participants on their ability to distinguish native and non-native phones. As expected Bohn [2017], Kuhl et al. [2005], participants were better at discriminating native sounds than non-native ones (across phone pairs: $p < 10^{-18}$, Figure 4-A). Second, applying the same test to our self-supervised French and English models shows that, like humans, models best discriminate sounds from their 'native' language (i.e., the French model better distinguishes French stimuli than English ones, across phone pairs, and vice versa: $p < 0.05$). Finally, the random and acoustic models obtain the worst ABX accuracy. These results confirm that 600 h of self-supervised learning suffices for Wav2Vec 2.0 to learn language-specific representations (Figure 4-B).

Wav2Vec 2.0 and the brain learn language specific representations. Next, we compare the Neural Predictivity scores of random, non-speech, non-native and native models (Figure 4-C D). First, our results show that the non-speech model attains higher R scores than the random model (on average across voxels, $\Delta R = 0.006$, p=10^{-31}) confirming the importance of learning to generate brain-like representations. Second, non-native models attain higher R scores than the non-speech model ($\Delta R = 0.002, p = 10^{-9}$), confirming that Wav2Vec 2.0 learns speech-specific representations of sounds when trained on speech. Finally, the native model attains higher R scores than non-native models ($\Delta R = 0.002, p = 10^{-15}$). The specialization for native sounds shared between models and the brain appears to be mainly localized around the superior temporal sulcus and the middle temporal gyrus (Figure 4-D), whereas higher-level regions like IFG and the angular gyrus only show more general specializations (i.e., speech rather than native-language sounds).

4 Discussion

Human infants acquire language with little to no supervision. A few hundred hours of speech suffices for young brains to learn to discretize phonemes, segment morphemes, and assemble words in the language(s) of their social group Dupoux [2018]. Here, we test whether self-supervised learning on a limited amount of speech suffices to yield a model functionally equivalent to speech perception in the human brain. We train several variants of Wav2Vec 2.0 Baevski et al. [2020] with three curated datasets of French, English, and Mandarin, and compare their activations to those of a large group of French, English, and Mandarin speakers recorded with fMRI while passively listening to audio stories. Our results show that this self-supervised model learns (1) representations that linearly map onto a remarkably distributed set of cortical regions (Figure 2), (2) a hierarchy that aligns with the cortical hierarchy (Figure 3), and (3) features specific to the language of the participants (Figure 4).

These results extend recent findings on the similarities between brain responses to speech and deep learning models trained with biologically-implausible objectives and data. First, fMRI Kell et al. [2018], Millet and King [2021], Thompson et al. [2021], electroencephalography Huang et al. [2018], and multi- or single-unit responses to sounds Koumura et al. [2019], Begus et al. [2022] have been shown to be linearly predicted by the activations of deep convolutional networks trained on *supervised* auditory tasks. For example, Millet and King [2021] showed that a supervised model, DeepSpeech 2.0 Amodei et al. [2016], better accounted for brain responses to speech in 102 individuals when it was trained on speech recognition rather than auditory scene classification. Similarly, Kell et al. [2018] showed that eight participants listening to brief speech and non-speech sounds demonstrated fMRI responses in the temporal lobe that aligned with those of a deep convolutional neural network trained on a dual auditory classification task. Our results, based on up to 50 times more fMRI recordings of the entire cortex, not only show that such representational similarities hold with a self-supervised objective, but reveal the remarkably similar hierarchical organisation of speech processing between these two systems Lerner et al. [2011], Berezutskaya et al. [2017], Caucheteux et al. [2021b,a].

Second, a growing series of MEG Toneva and Wehbe [2019], Caucheteux and King [2022], fMRI Qian et al. [2016], Pereira et al. [2018], Schwartz et al. [2019], Antonello et al. [2021] and electrophysiology studies Schrimpf et al. [2021], Goldstein et al. [2022] show that text-based language models trained on very large corpora generate brain-like representations too. While these results suggest elements of convergence between language models and the brain, they remain largely biologically implausible: not only are these algorithms pre-equipped with abstract linguistic units such as characters and words, but they are trained on corpora that no one would ever be able to read in their lifetime. In contrast, Wav2Vec 2.0 is here trained with raw speech waveforms, in a quantity that is comparable to what human infants require for early language acquisition Dupoux [2018].

The present study reveals the neural hierarchies and specializations underlying speech comprehension in the brain with remarkable clarity. First, this hierarchy is aligned with the anatomy: *e.g.* the superior temporal sulcus and the temporal pole are known to project to the ventral and dorsal part of the inferofrontal gyrus, respectively Petkov et al. [2015]. Second, the identification of functional gradients within the prefrontal cortex, and down to the motor areas typically associated with larynx and mouth control Dichter et al. [2018] reinforces the relevance of motor processes to speech perception Kellis et al. [2010], Mugler et al. [2014], Shamma et al. [2021]. Finally, the existence of multiple levels of representations around the inferofrontal cortex is consistent with the idea that Broca's area may be responsible for merging linguistic units Chomsky [2000], Friederici [1999], Hagoort [2005], Poeppel et al. [2012]. It should be noted, however, that our results aggregate a large cohort of individuals which could mask a more modular organization at the individual level.

Still, several major gaps remain between self-supervised speech models like Wav2Vec 2.0 and the brain. First, its transformer layers are not temporally constrained: each layer can access all elements within the contextual window. This differs from the necessarily recurrent nature of processing in the brain. Second, Wav2Vec 2.0 shows differences with humans in recent behavioral studies: it shows a heightened sensitivity to band-pass filtering and an under-reliance on fine temporal structures Weerts et al. [2021]. It also fails to predict categorical effects on perception Millet et al. [2021]. Third, recent studies show that Wav2Vec 2.0 encodes significantly less semantic information than text-based models Pasad et al. [2021]. These limitations may be due to the time scales of Wav2Vec 2.0 which, unlike humans, learns very short-term representations of speech. Overall, given that the human brain remains the best known system for speech processing, our results highlight the importance of systematically evaluating self-supervised models on their convergence to human-like speech representations.

The complexity of the human brain is often thought to be incompatible with a simple theory: “Even if there were enough data available about the contents of each brain area, there probably would not be a ready set of equations to describe them, their relationships, and the ways they change over time” Gallant [2013]. By showing how the equations of self-supervised learning give rise to brain-like processes, this work is an important challenge to this view.

Acknowledgements

This research was supported by the École Doctorale Frontières du Vivant (FdV) – Programme Bettencourt, and by French Agence Nationale de la Recherche grants ANR-17-CE28-0009 (GEOMPHON), ANR-11-IDFI-023 (IIFR), ANR-11-IDEX-0005 (USPC), ANR-10-LABX-0083 (EFL), ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute to JM, PO, YB, ED, JRK for their work at PSL. This work was performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011012415, 2022-AD011012415R1).

References

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34, 2021.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *LREC*, 2020.

Alexei Baevski, H. Zhou, Abdel rahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477, 2020.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Gasper Begus, Alan Zhou, and Christina Zhao. Encoding of speech in convolutional layers and the brain stem based on language experience. *bioRxiv*, 2022.

Yoav Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 72(4):405–416, 2010.

Julia Berezutskaya, Zachary V Freudenburg, Umut Güçlü, Marcel AJ van Gerven, and Nick F Ramsey. Neural tuning to low-level features of speech throughout the perisylvian cortex. *Journal of Neuroscience*, 37(33):7906–7920, 2017.

Ocke-Schwen Bohn. Cross-language and second language speech perception. *The handbook of psycholinguistics*, pages 213–239, 2017.

Lasse Borgholt, Jakob Drachmann Havtorn, Joakim Edin, Lars Maaløe, and Christian Igel. A brief overview of unsupervised neural speech representation learning. *arXiv preprint arXiv:2203.01829*, 2022.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, 2022.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Long-range and hierarchical language predictions in brains and algorithms. *arXiv:2111.14232 [cs, q-bio]*, November 2021a. arXiv: 2111.14232.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3635–3644, Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics.

Noam Chomsky. Linguistics and brain science. *Image, language, brain*, pages 13–28, 2000.

Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

Beijing Magic Data Technology Co. Magic data chinese mandarin conversational speech. `http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101`, 2019.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.06.010.

Benjamin K Dichter, Jonathan D Breshears, Matthew K Leonard, and Edward F Chang. The control of vocal pitch in human laryngeal motor cortex. *Cell*, 174(1):21–31, 2018.

Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Angela D Friederici. The neurobiology of language comprehension. In *Language comprehension: A biological perspective*, pages 265–304. Springer, 1999.

Jack Gallant. in 'reading minds'. *Nature*, 502(7472):428, 2013.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Alex Graves. Connectionist temporal classification. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 61–93. Springer, 2012.

Peter Hagoort. On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9): 416–423, 2005.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Nicholas Huang, Malcolm Slaney, and Mounya Elhilali. Connecting deep neural networks to physical, perceptual, and electrophysiological auditory signals. *Frontiers in neuroscience*, 12:532, 2018.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, April 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature17637.

Alexander JE Kell and Josh H McDermott. Deep neural network models of sensory systems: windows onto the role of task constraints. *Current opinion in neurobiology*, 55:121–132, 2019.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Spencer Kellis, Kai Miller, Kyle Thomson, Richard Brown, Paul House, and Bradley Greger. Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, 7(5):056007, 2010.

Tim C Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep neural networks in computational neuroscience. *BioRxiv*, page 133504, 2018.

Takuya Koumura, Hiroki Terashima, and Shigeto Furukawa. Cascaded tuning to amplitude modulation for natural sound recognition. *Journal of Neuroscience*, 39(28):5517–5533, 2019.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

Patricia K Kuhl, Barbara T Conboy, Denise Padden, Tobey Nelson, and Jessica Pruitt. Early speech perception and later language development: Implications for the" critical period". *Language learning and development*, 1(3-4):237–264, 2005.

Y. Lerner, C. J. Honey, L. J. Silbert, and U. Hasson. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, 31(8):2906–2915, February 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3684-10.2011.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John T Hale. Le petit prince: A multilingual fmri corpus using ecological stimuli. *bioRxiv*, 2021.

Juliette Millet and Ewan Dunbar. Do self-supervised speech models develop human-like perception biases? *AAAI 2022 Workshop, Self-Supervised Learning for Audio and Speech Processing*, 2022.

Juliette Millet and Jean-Remi King. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv preprint arXiv:2103.01032*, 2021.

Juliette Millet, Ioana Chitoran, and Ewan Dunbar. Predicting non-native speech perception using the perceptual assimilation model and state-of-the-art acoustic models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 661–673, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.51. URL `https://aclanthology.org/2021.conll-1.51`.

Emily M Mugler, James L Patton, Robert D Flint, Zachary A Wright, Stephan U Schuele, Joshua Rosenow, Jerry J Shih, Dean J Krusienski, and Marc W Slutzky. Direct classification of all american english phonemes using signals from functional speech motor cortex. *Journal of neural engineering*, 11(3):035015, 2014.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. Narratives: fmri data for evaluating models of naturalistic language comprehension. preprint. *Neuroscience, December*, pages 2020–06, 2020.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. *arXiv preprint arXiv:2107.04734*, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018.

Christopher I. Petkov, Yukiko Kikuchi, Alice E. Milne, Mortimer Mishkin, Josef P. Rauschecker, and Nikos K. Logothetis. Different forms of effective connectivity in primate frontotemporal pathways. *Nature Communications*, 6(1):6000, January 2015. ISSN 2041-1723. doi: 10.1038/ncomms7000. Number: 1 Publisher: Nature Publishing Group.

David Poeppel, Karen Emmorey, Gregory Hickok, and Liina Pylkkänen. Towards a new neurobiology of language. *Journal of Neuroscience*, 32(41):14125–14131, 2012.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. Bridging lstm architecture and the neural dynamics during reading. *arXiv preprint arXiv:1604.06635*, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Thomas Schatz. *ABX-discriminability measures and applications*. PhD thesis, Université Paris 6 (UPMC), 2016.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems*, 32, 2019.

Shihab Shamma, Prachi Patel, Shoutik Mukherjee, Guilhem Marion, Bahar Khalighinejad, Cong Han, Jose Herrero, Stephan Bickel, Ashesh Mehta, and Nima Mesgarani. Learning speech production and perception through sensorimotor interactions. *Cerebral cortex communications*, 2(1):tgaa091, 2021.

Jessica AF Thompson, Yoshua Bengio, Elia Formisano, and Marc Schönwiesner. Training neural networks to recognize speech increased their correspondence to the human auditory pathway but did not yield a shared hierarchy of acoustic features. *bioRxiv*, 2021.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv:1905.11833 [cs, q-bio]*, November 2019. arXiv: 1905.11833.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

Lotte Weerts, Stuart Rosen, Claudia Clopath, and Dan FM Goodman. The psychometrics of automatic speech recognition. *bioRxiv*, 2021.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE, 2021.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3
   (b) Did you describe the limitations of your work? [Yes] See Section 4
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have, and since we use results of human experiments that were previously made, and that respect these review guidelines, we conform to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results
   (b) Did you include complete proofs of all theoretical results? [N/A] No theoretical results
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code will be provided upon request. Otherwise for the rest, see Section 2.2 and 2.1.3
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2.1.3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] This is the case for all our figures, see Figure 2, 3 and 4
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The models we train are not new: the original papers already give that type of information.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 2.2, 2.1.3, and 2.4.
   (b) Did you mention the license of the assets? [No] They are all under CC0 for the English and French training audio data, the Perceptimatic Benchmark, The Little Prince dataset and the Narratives dataset. Magic data is under CC BY-NC-ND 4.0, Audioset is under CC BY 4.0 for the audio recordings, and CC BY-SA 4.0 for the ontology, Narratives' preprocessing code is under GNU GENERAL PUBLIC LICENSE and fairseq code is under MIT License.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] The code will be given upon request
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 2.2 and 2.4. More details on the way they gave their consent can be seen in the original papers/websites that provided the experimental results.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data were anonymized before we used them.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not conduct the experiments ourselves
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not conduct the experiments ourselves
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not conduct the experiments ourselves

A Appendix

A.1 Self-supervised loss formula

Wav2Vec 2.0, when trained in a self-supervised way, uses a loss (L) which is the weighted combination of two losses: one diversity loss (L_d), which pushes the quantization module to contain representations that are as diverse as possible, and one Contrastive Predictive Coding loss (L_m), which pushes the model to choose, from the context network output c, the right quantized representation (q) of some masked input, among other possible representations. L_m has the following formula, for some masked time step t:

$$\mathcal{L}_m = -\log \frac{\exp\left(\operatorname{sim}\left(\mathbf{c}_t, \mathbf{q}_t\right)/\kappa\right)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp\left(\operatorname{sim}\left(\mathbf{c}_t, \tilde{\mathbf{q}}\right)/\kappa\right)} \tag{2}$$

with $\operatorname{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}/\|\mathbf{a}\|\|\mathbf{b}\|$, κ the temperature, which is constant during training, Q_t the set of $K+1$ quantized candidate the model has to choose from, including the right one, i.e. q_t.

L_d is included to encourage the equal use of the V possible entries of each of the G codebooks of the quantization module. The goal is to maximize the entropy of the averaged softmax distribution over the codebook entries for each codebook $\bar{p}_g$, across a set of utterances:

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H\left(\bar{p}_g\right) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v} \tag{3}$$

A.2 Supervised loss formula

When trained in a supervised way, Wav2Vec 2.0 is trained to optimise a Connectionist Temporal Classification loss parameterized over θ:

$$\operatorname{argmin}_\theta \; -\log \sum_{a \in a_{U,V}} \prod_{t=1}^{d_t} p_{\mathrm{CTC}}\left(a_t \mid m_\theta(U)\right) \quad , \tag{4}$$

where $m_\theta(U) \in \mathbb{R}^{d_\tau \times d_v}$ are the probabilistic predictions of the model at each τ time sample given the input raw waveform $U \in \mathbb{R}^{d_\tau \times d_u}$, $V \in \mathbb{R}^{d_t \times d_v}$ are the true transcriptions of U, and $a_{U,V}$ is the set of all possible alignments between U and V.

A.3 Preprocessing of the model's activations

The activations of the network $X \in \mathbb{R}^{d_{\hat{t}} \times d_x}$ are first normalized to be between $[0, 1]$ for each listening session. Then, we use nistats Abraham et al. [2014] `compute_regressor` function with the 'glover' model to temporally convolve ($h \in \mathbb{R}^{d_{\hat{t}}}$) and temporally down-sample (using $g : \mathbb{R}^{d_{\hat{t}}} \to \mathbb{R}^{d_t}$) each artificial neuron j:

$$\hat{x}^{(j)} = g\Big(x^{(j)} * h\Big) \quad . \tag{5}$$

A.4 Penalized linear model - Ridge regression

For each split s, we fit an ℓ_2-penalized linear model $V \in \mathbb{R}^{d_x \times d_z}$ trained to predict the transformed BOLD time series from the model activations for each dimension independently. The formula of the optimization is the following:

$$\operatorname{argmin}_V \sum_{i \in \mathrm{train}_s} (V^\top \hat{X}_i - y_i)^2 + \lambda \|V\|^2 \quad . \tag{6}$$