Encoding of speech in convolutional layers and the brain stem based on language experience

Gašper Beguš[a], Alan Zhou[a], T. Christina Zhao[b,c]

[a] *Department of Linguistics, University of California, Berkeley, United States*
[b] *Institute for Learning and Brain Sciences, University of Washington, United States*
[c] *Department of Speech and Hearing Sciences, University of Washington, United States*

Abstract

Comparing artificial neural networks (ANNs) with outputs of brain imaging techniques has recently seen substantial advances in (computer) vision and text-based language models. Here, we propose a framework to compare biological and artificial neural computations of spoken language representations and propose several new challenges to this paradigm. Using a technique proposed by Beguš and Zhou (2021b), we can analyze encoding of any acoustic property in intermediate convolutional layers of an artificial neural network. This allows us to test similarities in speech encoding between the brain and artificial neural networks in a way that is more interpretable than the majority of existing proposals that focus on correlations and supervised models. We introduce fully unsupervised deep generative models (the Generative Adversarial Network architecture) trained on raw speech to the brain-and-ANN-comparison paradigm, which enable testing of both the production and perception principles in human speech. We present a framework that parallels electrophysiological experiments measuring complex Auditory Brainstem Response (cABR) in human brain with intermediate layers in deep convolutional networks. We compared peak latency in cABR relative to the stimulus in the brain stem experiment, and in intermediate convolutional layers relative to the input/output in deep convolutional networks. We also examined and compared the effect of prior language exposure on the peak latency in cABR, and in intermediate convolutional layers of a phonetic property. Specifically, the phonetic property (i.e., VOT =10 ms) is perceived differently by English vs. Spanish speakers as voiced (e.g. [ba]) vs voiceless (e.g. [pa]). Critically, the cABR peak latency to the VOT phonetic property is different between English and Spanish speakers, and peak latency in intermediate convolutional layers is different between English-trained and Spanish-trained computational models. Substantial similarities in peak latency encoding between the human brain and intermediate convolutional networks emerge based on results from eight trained networks (including a replication experiment). The proposed technique can be used to compare encoding between the human brain and intermediate convolutional layers for any acoustic property.

Keywords: Auditory brainstem response, interpretable deep learning, speech, unsupervised learning

URL: begus@berkeley.edu (Gašper Beguš), azhou314@berkeley.edu (Alan Zhou), zhaotc@uw.edu (T. Christina Zhao)

1. Introduction

Many aspects of artificial neural networks (ANNs) are biologically inspired and have equivalents in the human brain, but several properties are nevertheless biologically implausible (Pulvermüller et al., 2021; Bengio et al., 2016; Whittington and Bogacz, 2019; Marblestone et al., 2016). Among the architectures highly influenced by brain processing in the visual domain are convolutional neural networks (Fukushima, 1980; LeCun et al., 1989; Yamins and DiCarlo, 2016; Kell and McDermott, 2019; Lindsay, 2021; la Tour et al., 2021). Despite the fact that many aspects of current ANNs lack biological plausibility, it is nevertheless reasonable to compare computations and representations in deep neural network and the brain. Such work has twofold implications. On the one hand, the comparison has the potential to shed light on how ANNs encode representations internally relative to the brain and how learning biases in humans and ANNs differ. On the other hand, computational models allow us to simulate brain processes (such as speech) and test hypotheses that are not possible to test in the human brain. Such simulations can bring insights for how language gets acquired and encoded in the brain. For example, we can test what properties of speech (both in terms of behavioral and neural data) emerge when models have no articulatory biases compared to models with articulatory information, or when models have no language-specific mechanisms. Such questions can help us better understand which properties of language are domain specific vs. domain general, and which properties emerge from articulatory factors (see the discussion in Section 1.3). Such experiments that cannot be conducted on human subjects necessarily require computational simulations.

The majority of work comparing the brain and ANNs is performed on the visual domain with substantially fewer studies comparing ANNs to brain responses to linguistic stimuli. Most existing comparison studies in the linguistic domain focus on text-trained models and supervised models, and focus on correlations (see Section 1.1). Here, we outline a technique that parallels neural encoding of specific acoustic phonetic features by examining ANN models trained on raw speech in a fully unsupervised manner. We introduce the GAN architecture (Goodfellow et al., 2014) to the brain-ANN comparison literature.

GANs are uniquely appropriate for modeling speech acquisition (Beguš, 2020b, 2021a). Crucially, GANs need to learn to generate output from noise by imitation. The main characteristic of the architecture are two networks, the Generator and the Discriminator, that are trained in a minimax game (Goodfellow et al., 2014), in which the Discriminator attempts to distinguish real data and outputs from the Generator, and the Generator learns to generate realistic output given only feedback from the Discriminator (summary in Figure 2). It has been shown that this process results in the ability to encode linguistic information (e.g. lexical and sublexical representations) into raw speech in a fully unsupervised manner (Beguš, 2021a) as well as in the ability to learn highly complex morphophonological rules (Beguš, 2021b) both locally and non-locally (Beguš, 2021c). In other words, linguistically meaningful representations (such as words, prefixes) self-emerge in the GAN architecture when the models are trained on raw speech. Evidence for several hallmarks of symbolic-like representations emerge in GANs: discretized (disentagled) representations, a causal relationship between the latent space and generated outputs, and near-categoricity of desired outputs (Beguš, 2021a,b). Crucially, GANs do not simply replicate input data (as is the case for autoencoders), but generate innovative and interpretable outputs, which mimics one of the more prominent features of language: productivity (Piantadosi and Fedorenko, 2017).

1.1. Prior work

As already mentioned, a substantial amount of work exists on paralleling brain imaging with artificial neural networks in the visual domain (Cadieu et al., 2014; Güçlü and van Gerven, 2015;

Yamins and DiCarlo, 2016; Cichy et al., 2016; Greene and Hansen, 2018; Eickenberg et al., 2017; Storrs and Kriegeskorte, 2019; Lindsay, 2021) and relatively fewer works exist in the language or in the speech domain (for a text-based language model, see Jain and Huth 2018; Jat et al. 2019; Schrimpf et al. 2020). Kell et al. (2018) and Millet and King (2021) parallel fMRI recordings with supervised ASR models that are trained on spectrograms. The comparisons in Kell et al. (2018) and Millet and King (2021) reveal parallels in neural encoding between ANNs and the brain, but are based on a linear regression estimates between the two sets of signals, focuses on correlations, and do not directly compare individual acoustic properties without linear transformations. Huang et al. (2018) examine a measurement of surprisal in an supervised CNN classifier, and correlate the metric to an EEG signal reduced in dimensionality. Donhauser and Baillet (2020) train a predictive ANN model and use it to quantify the brain's response to surprisal during speech processing. Koumura et al. (2019) focus on amplitude modulation of auditory stimuli (not only of speech). Their model is trained on raw waveforms, but the analysis focuses on individual units in deep convolutional networks. They analyze synchrony and average activity for each unit and analyze them across convolutional layers. All their models are fully supervised classifiers (thus modeling only perception) and do not focus on linguistically meaningful representations, but on acoustic phonetic properties and audition in general. Smith et al. (2021b,a) argue for parallels in human binaural detection and deep neural networks (VAEs). They model pure tones rather than speech and focus on binaural detection. Khatami and Escabí (2020) operate with hierarchical spiking neural networks on cochleograms using supervised training. All these proposals focus on correlations or similarity scores. The speech datasets in all papers except in Millet and King (2021) are limited to one language—English from TIMIT or from other corpora. Saddler et al. (2021) compare supervised deep convolutional networks for F0 classification with models of the auditory nerve, but not with actual brain imaging data. All these frameworks use supervised classification networks for their comparison. Below, we outline how our model differs from these existing proposals.

1.2. Goals & new challenges

This paper proposes some crucial new approaches and guidelines to the comparison of learned spoken language representations in deep neural networks and the brain. First, we compare brain data to fully unsupervised models where linguistically meaningful representations need to self-emerge. Unsupervised learning resembles human speech acquisition more closely than supervised models trained for automatic speech recognition or acoustic scene classification tasks. Rather than on pre-trained models, we train the networks on controlled data which allows for more interpretable results and a more controlled experimental setting.

Second, our models and visualization techniques capture both the production and perception component in human speech (equivalent to encoding and decoding; for a discussion of the two concepts in cognitive science, see Kriegeskorte and Douglas 2019), while most existing proposals exclusively focus on the perception component. We perform comparison between brain and ANN data from both the Generator network that simulates speech production (synthesis, decoding) and the Discriminator network that simulates speech perception (classification, encoding). For modeling the production element, we propose a procedure for comparing ANNs with the brain where the model's internal elements (latent space) are chosen such that the model's generated output and the stimulus in the brain imaging experiment are maximally similar (using a technique in Lipton and Tripathi 2017; Keyes et al. 2020). This is the opposite direction of a related proposal in Norman-Haignere and McDermott (2018). For modeling the perception element, we feed the Discriminator network outputs of the Generator that are forced to resemble the stimulus. The production and perception in human speech are highly interconnected (Vihman, 2015), which is why modeling both principles is highly desired when comparing brains and ANNs.

Third, instead of focusing on correlations between some metric in brain imaging experiments and some other metric in deep neural networks, we focus on comparing interpretable features across the two systems. In this paper, we focus on peak latency in both the cABR and in deep convolutional neural networks. This is an interpretable acoustic property, is directly comparable, and requires no computation of correlations or any transformations/regressions between signals. Comparing acoustic properties rather than correlations is both more interpretable and less problematic: correlations can arise even on untrained models and are generally problematic to analyze and interpret.

Fourth, most of the existing proposals focus on correlating brain responses and outputs of neural networks in a single language, which primarily models neural encoding of the acoustic speech signal and does not specifically model any phonological contrasts across languages. By training the networks on two languages with a different encoding of a phonetic property (as confirmed by brain experiments), we not only test the encoding of acoustics, but we test specific phonetic features that constitute phonological contrasts: the distinction between voiceless stops (such as [t]) and voiced stops (such as [d]) in English and Spanish. How *phonological* (meaning-distinguishing) contrasts are encoded in the brain Zhao and Kuhl (2018) and in deep neural network trained on speech Beguš (2020b, 2021b,a,c) can yield new information on encoding of linguistically meaningful units across the two systems.

Fifth, we propose a technique to compare EEG signals with deep neural networks (for a comparison between EEG signals and ANNs in the visual domain, see Greene and Hansen 2018; for speech, see Huang et al. 2018). Unlike other brain imaging techniques (e.g. fMRI or ECoG), EEG is minimally invasive while providing high temporal resolution, which is crucial for examining speech encoding as they are temporally dynamic signals. This should allow a large-scale comparison between deep neural networks and the brain not only for those phonetic properties investigated in this paper, but any other acoustic property.

Finally, we aim to show that earlier layers in deep neural networks correspond to earlier processing of speech in the brain. For this reason, we focus on the complex auditory brainstem response (cABR), a potential that can robustly reflect sensory encoding of auditory signals in early stages of auditory processing (Skoe and Kraus, 2010). We parallel cABR signal to the second to last convolutional layer in deep convolutional networks when modelling production, and in the first layer in networks when modelling perception. Comparing cABRs and deep networks is, to our knowledge, new in the paradigm of comparing deep learning and the brain. Unlike other imaging techniques (such as fMRI or ECoG), cABR is one of the few brain imaging techniques that allows recording of the brainstem regions and captures the earliest stages of speech processing. Recent evidence suggests that several acoustic properties that result in phonological contrasts are encoded already in the brain stem (Zhao and Kuhl, 2018; Zhao et al., 2019).

To achieve these goals, we compare outputs of the cABR experiment from Zhao and Kuhl (2018) to the fourth/first convolutional layer (out of five total layers) in deep convolutional networks. The networks are trained in a Generative Adversarial Network framework (Goodfellow et al., 2014), where the Generator network learns to produce speech from some random latent distribution and the Discriminator learns to distinguish real from generated samples. In other words, the Generator needs to learn to produce speech-like units in a fully unsupervised way — it never actually accesses real data, but rather needs to trick another network by producing real-looking data outputs. This unsupervised learning process based on imitation, where the networks learn to generate data from noise based only on unlabeled data, more closely resembles language acquisition than competing proposals. We train the networks on sound sequences that closely resemble the stimulus in the cABR experiments and are sliced from two corpora — one on English (TIMIT; Garofolo et al. 1993) and one on Spanish (DIMEx; Pineda et al. 2004), simulating the monolingual English and Spanish subjects in the cABR experiment.

We then propose a new technique comparing brain imaging and deep neural networks. To test internal representations in the artificial neural network that simulates the production of speech, we force the Generator to output sequences of speech sounds that are as similar to the stimulus used in the cABR experiment as possible. To test internal representations of networks that simulates perception of speech, we feed the generated outputs that closely resemble the stimulus to the Discriminator network. Using the visualization techniques proposed in Beguš and Zhou (2021b), we can compare any acoustic property of speech in the generated output/input and the previous internal convolutional layers in either the Generator (simulating speech production) and the Discriminator network (simulating speech perception). The comparison is then performed between (i) the generated outputs in deep neural networks, and the second-to-last convolutional layer in the Generator, and the first convolutional layer in the Discriminator and (ii) the stimulus played to subjects during the experiment and averaged cABR recording in the brain stem. We argue that this technique yields interpretable results — we can take any acoustic property and compare its encoding in the brain and in the artificial neural networks. To test how language experience alters representations in the brain and in artificial neural networks, we perform the comparisons on monolingual subjects of two languages in the neuroimaging experiment and deep learning models trained on two languages.

The results of this technique presented in this paper suggest that peak latency differs in similar ways in the brain stem and in deep convolutional neural networks depending on which languages subjects/models are exposed to. To avoid idiosyncrasies in the models, we replicate the experiment. Results are consistent across all 8 sets of generated outputs from 4 independently trained models.

1.3. Limitations of comparison between brains and deep neural networks

Comparing representations and computations in the human brain and deep learning models is a complex task. The goal of this paper is not to argue that human speech processing operates exactly as in deep convolutional networks (for problems with such an approach, see Guest and Martin 2021). We do, however, show, that computations and encodings are similar. By focusing on interpretable comparison (rather than on correlations) and by focusing on internal representations rather than only on behavioral data (see discussion in Guest and Martin 2021), we argue that similarities in both representations and computations exist between brains and deep convolutional layers. These similarities open up possibilities for modeling work in order to gain insights both for how humans acquire and process speech as well as for how deep learning models learn internal representations.

For example, our models are closer to reality than most existing models because the learning is fully unsupervised and the models are trained on raw speech which requires no preabstraction or feature extraction (Beguš, 2020b, 2021b). The models, however, still feature several unrealistic properties. First, humans do not learn language exclusively from audio data. Second, the models contain no articulatory information, while we know that humans produce speech with articulators. While these limitations are undesired because they make models less realistic, they can also be advantageous from a cognitive modeling perspective. A long-standing debate in linguistics and speech science concerns whether typological tendencies in speech patterns across the world's languages result from articulatory pressures and transmission of language in space and time, or result from cognitive biases (Kiparsky, 2006, 2008; Blevins, 2013; Beguš, 2019, 2020a). Another major debate in linguistics and cognitive science assesses which properties of language are domain-specific and innate and which can be explained by domain-general cognitive principles (Culbertson and Kirby, 2016). Modeling speech processing in deep neural networks that contain no articulatory information and no language-specific elements allow us to test what internal representations emerge in unsupervised deep neural network trained on speech without articulators and without any language-specific elements. This can provide information on what design properties of language are possible without articulators and without language-specific devices. In fact, using the techniques proposed in this paper, we can now test these hypotheses not only behaviorally, but can also compare internal representations in models without these properties and in the brain.

2. cABR Experiment

The complex auditory brainstem response (cABR) reflects the early sensory encoding of complex sounds along the auditory pathway and can be measured with a 3-electrode set up using EEG (Skoe and Kraus, 2010). The cABR generally contains an onset component, corresponding to transient changes in acoustics (e.g., stop consonant) as well as a frequency-following-response component (FFR), corresponding to periodic portions of the sound (e.g., tone, vowel). In recent decades, there has been a growing literature on characteristics of cABR. Few studies that focused on speech perception have demonstrated evidence in support for important speech perception phenomenon at the cABR level. For example, native Mandarin speakers demonstrated FFR that tracks the pitch of the lexical tones better than English speakers, demonstrating that the language experiential effect can be observed at the encoding stage (Bidelman et al., 2011). Further, the directional asymmetry phenomenon in speech perception was also observed in FFR to vowels Zhao et al. (2019). Lastly, the cABR and behavioral perception of stop consonants are highly correlated, demonstrating the cABR's behavioral relevance in speech perception. In addition, both the behavior and cABR are modulated by language background (Zhao and Kuhl, 2018).

The cABR data came from this previously published dataset from Zhao and Kuhl (2018).[1] The experiment measured the cABR when native English and Spanish subject listened to a synthesized syllable, which was identified as /ba/ by English speakers and /pa/ by Spanish speakers. Data from a total of 15 Spanish and 14 English monolingual speakers were included in the analysis.

2.1. The stimulus

The stimulus is a CV syllable with a vowel /a/. The bilabial stop consonant has a Voice-Onset-Time (VOT) of +10ms and was synthesized by Klatt synthesizer in Praat software (Boersma & Weenink, 2009). The syllable with 0ms VOT was first synthesized with a 2ms noise burst and vowel /a/. The fundamental frequency of the vowel /a/ began at 95Hz and ended at 90Hz. The silent gap (10ms) was then added after the initial noise burst to create syllables with the positive VOT. The waveform and spectrogram of the stimulus are shown in Fig 1i. The duration of the syllable is 100ms. Critically, monolingual English speakers identified the stimulus as /ba/ whereas native Spanish speakers identified the stimulus as /pa/, as reported in a previous behavioral experiment (Zhao and Kuhl, 2018). Individuals' cABR were calculated by averaging across all available trials after standard preprocessing and trial rejection. Averaged values are visualized in Figure 1g. Further, the group-level cABR can be visualized by averaging over all subjects. The monolingual English group and the native Spanish group are represented in Figure 1d.

2.2. cABR data acquisition

The details of the recording methods can be found in Zhao and Kuhl (2018). Specifically, the cABR reported here is recorded using a traditional set-up of 3-EEG channels (i.e., CZ electrode on a 10-20 system, ground electrode on the forehead and the reference electrode on the right earlobe; Skoe and Kraus 2010). Two blocks of recordings (3,000 trials per block) were completed for each participants where trials were alternating in polarities.

[1] Data can be accessed at `https://osf.io/6fwxd/`.

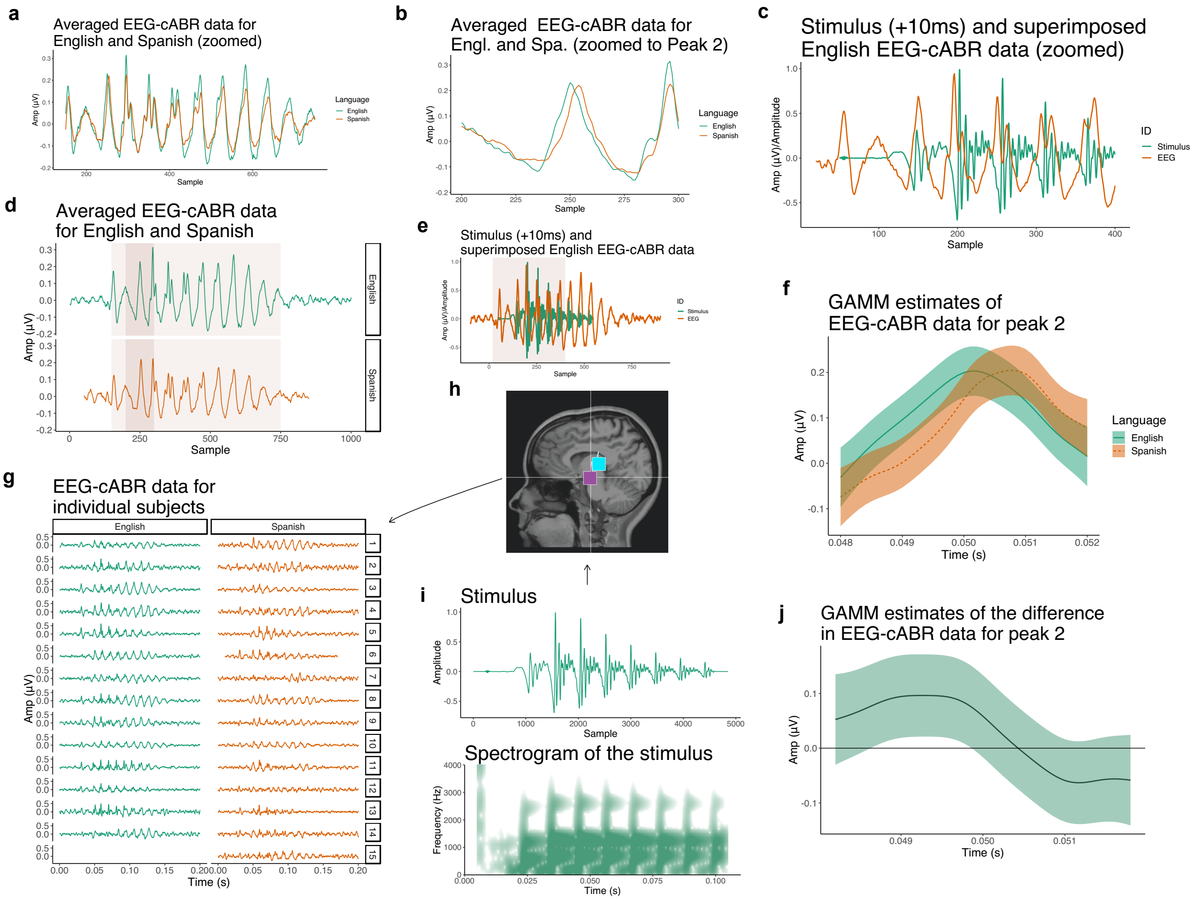

Figure 1: (**i**) Synthesized stimulus used in the cABR experiment with a spectrogram (0–4,000 Hz). This stimulus is also used in the computational experiments when the Generator is forced to output data with the objective to minimize the distance between generated data and the stimulus. The stimulus is played to subjects in the experiment. (**h**) The figure illustrates the dipole location of the onset peaks recorded during the cABR experiment for one speaker as localized in Zhao and Kuhl (2018) (magenta = peak 1, cyan = peak 2). The location suggests the recorded brain activity is indeed localized in the brain stem. Figure in (**g**) shows cABR recordings averaged for each subject across the 3,000 trials. (**d**) Individual subjects' recordings are averaged for each language (with shades that indicate which parts are zoomed in in the following figures). (**a**) Zoomed cABR data for English and Spanish showing that most peaks (with the exception of peak 2) are almost perfectly aligned across the two languages. (**b**) Zoomed peak 2 showing peak latency differences between English and Spanish. Figure in (**e**) superficially parallels the stimulus with the cABR data. The brain signal in the experiment is manually delayed relative to the stimulus; for illustration, we manually aligned the two time-series by approximately aligning the burst of the stimulus and the first peak of the cABR data. Shaded part indicates the area zoomed in (**c**). (**f**) Predicted values of a Generalized Additive Mixed Model with Amplitude in μV across time and the two languages (English vs. Spanish). For more details about the model, see Figure B.4 and Section 2.3. (**j**) Difference smooth between English and Spanish cABR data. The area in which the difference smooth's confidence interval do not cross zero indicates significant difference in the cABR signal between the two languages.

2.3. A new statistical analysis

Zhao and Kuhl (2018) show that peak latency timing differs significantly for peak 2 between English and Spanish subjects using independent t-test. To analyze data in a more interpretable fashion we fit the data averaged for each subject to Generalized Additive Mixed Models (GAMMs; Wood 2011) with the Amplitude of EEG-cABR in μV as the dependent variable and Language (treatment-coded with English as level) as parametric term, a smooth for time, by-language difference smooth for time, and by-speaker random smooths as well as correction for autocorrelation. The estimates of the model are in Table B.4. Despite random smooths, the models feature high degrees of autocorrelations. Significant difference does not arise for all windows of analysis likely due to correlation, but for a given window (from 240th to 260th sample), the difference smooth in Figure 1j suggest a significant difference in trajectory of the Amplitude between English and Spanish monolinguals in Peak 2 ($F = 2.70, p = 0.015$).

2.4. Results & interpretation of the cABR experiment

In summary, results from the cABR experiment demonstrated a robust effect of language background on the peak 2 latency of the cABR onset response. Particularly, the latency of peak 2, corresponding to the encoding of the onset of voicing, is significantly later in native Spanish speakers compared to the monolingual English speakers. Critically, the peak 2 latency was directly related to perception of the speech sound as shown in Zhao and Kuhl (2018). These suggest that the effect of language experience is reflected at very early stages of auditory processing, namely the auditory brainstem.

3. Computational Experiments

3.1. Model

We used the WaveGAN model (Donahue et al., 2019) for our computational experiments. WaveGAN is a 1D deep convolutional generative adversarial model that operates directly on the waveform itself. The Generator G uses 1D transpose convolutions to upsample from the latent space z, while the Discriminator D uses traditional 1D convolutions to predict the Wasserstein distance between the training distribution and the generated outputs $G(z)$ or real data x. The architecture is outlined in Figure 2.

WaveGAN itself does not contain any visualization techniques. For analyzing and visualization of intermediate convolutional layers, we use a technique proposed in Beguš and Zhou (2021b) (for the Generator) and Beguš and Zhou (2021a) for the classifier network (which has almost identical structure to the Discriminator). Beguš and Zhou (2021b,a) argue that averaging over feature maps after ReLU activation yields a highly interpretable time series data for each convolutional layer that summarize what acoustic properties are encoded at which layer.

For these experiments, we set z to be a 100-dimensional vector (following Donahue et al. 2019), which the Generator projects into a 2D tensor that is passed through 5 transpose convolutional layers, ending in an audio output of 16384 samples. The Discriminator similarly is composed of 5 (traditional) convolutional layers, with a hidden layer at the end that outputs the Wasserstein metric. The Discriminator also makes use of a process called phase shuffle (Donahue et al., 2019), which applies random perturbations to the phase of each layers' activations to prevent the Discriminator from accessing periodic artifacts characteristic of transpose convolutions.

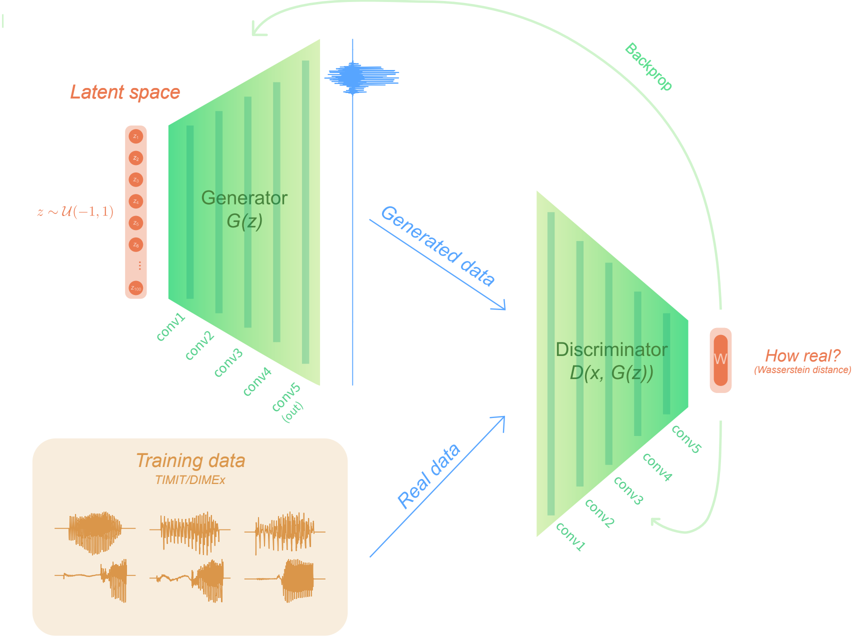

Figure 2: WaveGAN architecture Donahue et al. (2017) (based on Goodfellow et al. 2014; Radford et al. 2015) used in training. The training data was taken from TIMIT and DIMEx as described in Section 3.2.

	p	t	k	b	d	g
TIMIT	1018	1799	2112	1789	2530	673
DIMEx100	3015	1808	5558	1477	8023	478

Table 1: Counts of sequences beginning with each stop for each corpus.

3.2. Data

Spanish training data was taken from the DIMEx100 corpus (Pineda et al., 2004). This dataset consists of audio recordings of 5010 sentences in Mexican Spanish, recorded from 100 speakers mostly from around Mexico City. The dataset is balanced in gender and represents primarily the Mexico City variety of Spanish (Pineda et al., 2004). English training data was taken from the TIMIT speech corpus (Garofolo et al., 1993). The TIMIT dataset contains recordings of 6300 sentences of American English, spread across 8 dialects and 630 speakers (Garofolo et al., 1993).

For the purposes of training, we slice the first syllable from words that begin with a voiced or voiceless stop.[2] Specifically, we slice sequences of the form #TV, where # represents a word boundary, T represents a voiced or voiceless stop, and V represents a vowel. For both English and Spanish, the voiceless stops consist of [p, t, k] and the voiced stops consist of [b, d, g]. The number of sequences beginning with each stop in both datasets are shown in Table 1.

3.3. Training

We trained the DIMEx100 model for approximately 38,649 steps, after which model collapse was observed. To match the two models in the number of steps, we train the TIMIT model for

[2] For slicing, we modified code written by Sameer Arshad for the study in Beguš (2020b).

40,630 steps. To replicate the results and to control for idiosyncracies of individual models, we trained one additional TIMIT and one additional DIMEx100 model (for 41,818 and 39,417 steps, respectively).

3.4. Generating outputs that approximate the stimulus

In order to test the stimuli against the Generator network, we use latent vector recovery techniques (Lipton and Tripathi, 2017; Keyes et al., 2020) to find the latent variables that result in outputs closest to the stimuli. We then generate outputs using these latent variables and analyze each layer of the network given that latent space. This is a novel approach to paralleling representations in deep neural decoder networks and brain imaging outputs: the model's internal representations are chosen such that the generated output maximally resembles the stimulus in the brain experiment. Norman-Haignere and McDermott (2018) propose a somewhat similar procedure, where outputs of the brain experiments are paralleled with synthetic stimuli "designed to yield the same responses as the natural stimulus" (Norman-Haignere and McDermott, 2018). In our case, the directionality of forced input is reversed: we seek internal representations that result in maximal matching between the actual stimulus and the model's output.

We use gradient descent with stochastic clipping, as proposed in Lipton and Tripathi (2017), on the mean absolute error of the spectrogram of the stimulus and the spectrogram of the generated output as proposed in Keyes et al. (2020). We sample many random latent vectors uniformly for consistency, and optimize using the ADAM optimizer with learning rate of 1e−2, first moment decay of 0.9, and second moment decay of 0.99. We optimize for 10,000 steps, after which the majority of outputs converge. We adapt the objective function from Keyes et al. (2020). The function is listed below, where G is the generator network, $\mathcal{S}$ takes an audio signal to a spectrogram, and s is the target stimulus:

$$\min_{z^*} ||\mathcal{S}(s) - \mathcal{S}(G(z^*))||_1 \tag{1}$$

As the generator generates a fixed-length output, we must zero-pad the target stimulus before performing loss computations. Interestingly, while all training samples were simply right-padded to the desired dimension, we found that introducing varying amounts of left-padding had differing results in the quality of the generation. The DIMEx100 model, in particular, is extremely sensitive to the left pad, creating nonsense forced outputs with a left pad of 0 samples and creating much closer outputs with a left pad of 1000 samples. The TIMIT model is much less sensitive to the pad, and generates fairly close samples with a left pad of anything from 0 to 1000 samples. This difference may be due to differences in the slice distribution of the two corpora, but for the sake of consistency we used a left pad of 1000 samples for both models.

3.5. Procedure

The visualization technique in Beguš and Zhou (2021b,a) allows us to test acoustic representations of intermediate convolutional layers of both the Generator that mimics the production principle in speech and the Discriminator that mimics the perception principle. Here, we compare the outputs of the proposed technique to outputs of brain imaging experiments.

To analyze intermediate layers in deep convolutional layers and compare them to brain imaging outputs, we force the Generator to output #TV and #DV syllables that most closely resemble the stimulus used in the cABR experiment (Figure 1(i)), as described in Section 3.4.

The relationship in the brain between the stimulus played to the subjects in the cABR experiment and the amplitude of the cABR recording is modeled in this paper as the relationship between

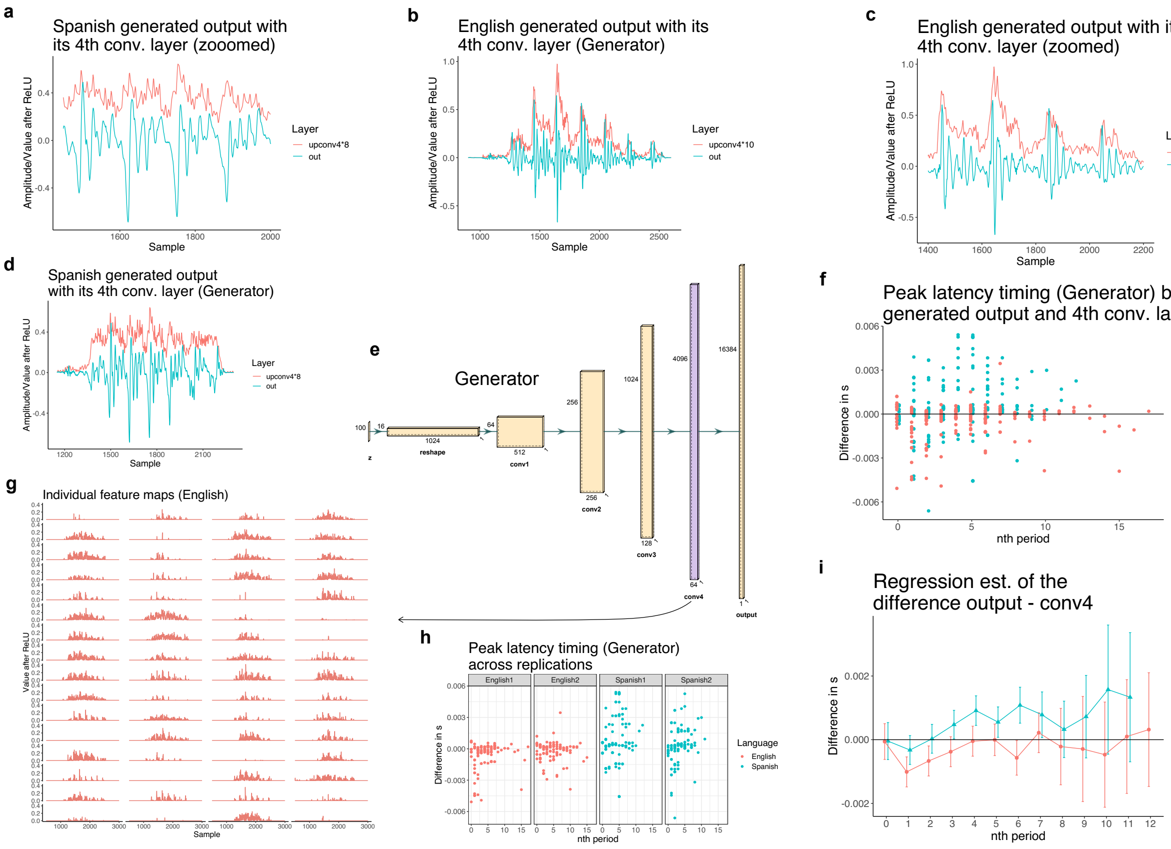

Figure 3: (**e**) The structure of the Generator network with five convolutional layers (Donahue et al., 2019). The fourth convolutional layer (Conv4; second to last) is color-coded with purple. (**g**) All 64 individual feature maps for a single output forced to closely resemble the stimulus from the fourth convolutional layer (Conv4) after ReLU (upsampled). (**d**) One Spanish output (in green) forced to resemble the stimulus with the corresponding values from the fourth convolutional layer (Conv4) averaged over all feature maps. The plot illustrates peak latency between output and Conv4 for the burst and each vocalic period. (**a**) A zoomed version of (e) focusing on four vocalic periods. (**b**) One English output (in green) forced to resemble the stimulus with the corresponding values from the fourth convolutional layer (Conv4) averaged over all feature maps. The plot illustrates peak latency between output and Conv4 for the burst and each vocalic period. (**c**) A zoomed version of (b) focusing on four vocalic periods. (**f**) Raw peak latency timing (output peak time - Conv4 peak time) for burst (=0) and each nth vocalic period across the two conditions (English vs. Spanish). Periods above the 12th period are rare and are discarded from the statistical analysis due to a small number of attestations. The data is pooled across the two replications. (**i**) Linear regression estimates for the peak latency timing between the two conditions (English vs. Spanish). Periods above the 12th period are discarded from the analysis due to a small number of attestations. The data is pooled across the two replications. (**h**) Raw peak latency timing across the replications (first and second replication) and two conditions (English and Spanish).

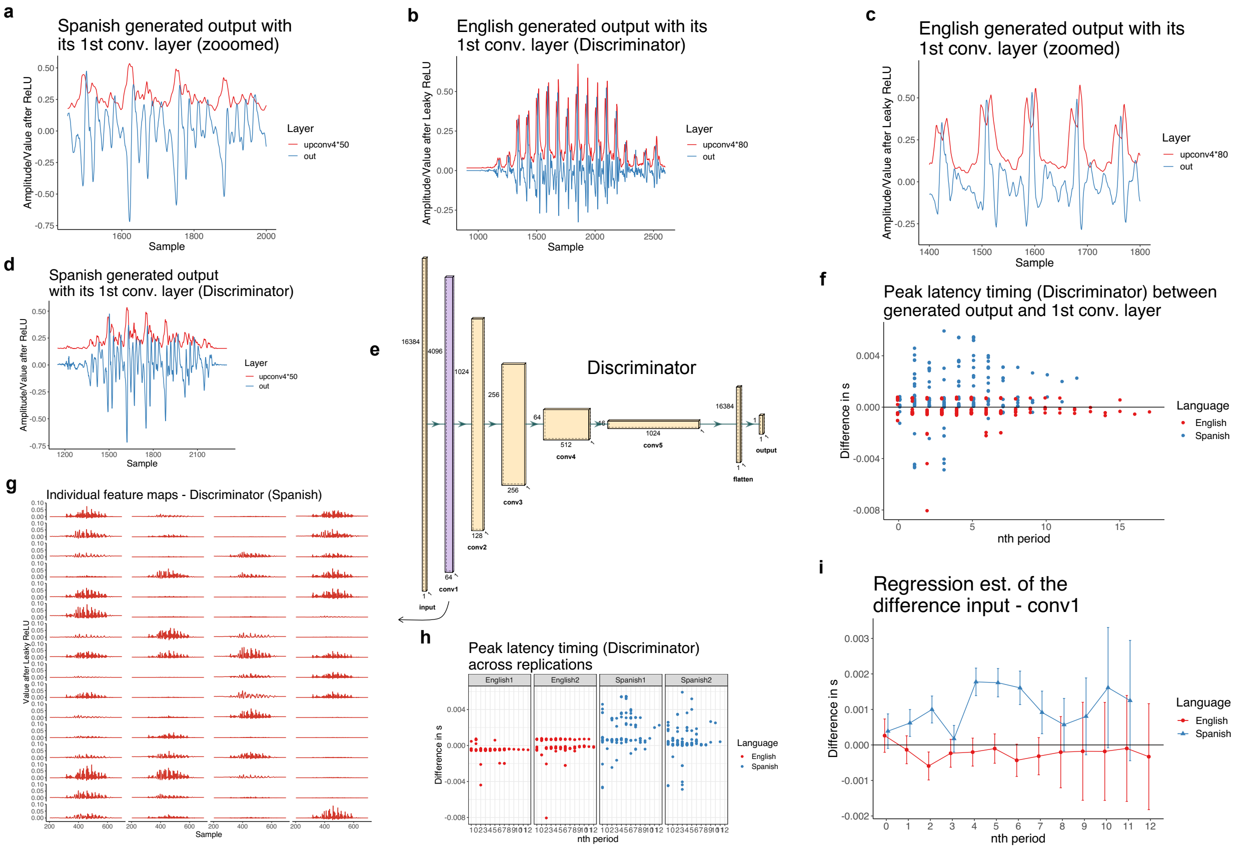

Figure 4: (**e**) The structure of the Discriminator network with five convolutional layers (Donahue et al., 2019). The first convolutional layer (Conv1; second to last) is color-coded with purple. (**g**) All 64 individual feature maps for a single input (the Generator's forced output) from the first convolutional layer (Conv1) after Leaky ReLU. (**d**) One Spanish input (in blue) from the Generator's forced output with the corresponding values from the first convolutional layer (Conv1) averaged over all feature maps. The plot illustrates peak latency between input and Conv1 for the burst and each vocalic period. (**a**) A zoomed version of (e) focusing on four vocalic periods. (**b**) One English input (in blue) from the Generator's forced output with the corresponding values from the first convolutional layer (Conv1) averaged over all feature maps. The plot illustrates peak latency between input and Conv1 for the burst and each vocalic period. (**c**) A zoomed version of (b) focusing on five vocalic periods. (**f**) Raw peak latency timing (input peak time - Conv1 peak time) for burst (=0) and each nth vocalic period across the two conditions (English vs. Spanish). Periods above the 12th period are rare and are discarded from the statistical analysis due to a small number of attestations. The data is pooled across the two replications. (**i**) Linear regression estimates for the peak latency timing between the two conditions (English vs. Spanish). Periods above the 12th period are discarded from the analysis due to a small number of attestations. The data is pooled across the two replications. (**h**) Raw peak latency timing across the replications (first and second replication) and two conditions (English and Spanish).

the generated output forced to resemble the stimulus and the fourth (second to last) convolutional layer in the Generator network.

As already mentioned, the Generator mimics the production aspect of speech. The cABR experiment, however tests encoding of a phonological contrast in the perception task. For this reason, we also test the relationship between the input to the Discriminator and its corresponding first convolutional layer as the Discriminator mimics the perception aspect of speech. The input to the Discriminator are the Generator's outputs forced to resemble the stimulus (according to Section 3.4).

The fourth and first convolutional layers, respectively, are analyzed according to Beguš and Zhou (2021b,a) by averaging over all feature maps after ReLU or Leaky ReLU activations.[3] This results in a time series t for each Convolutional layer as in (2) from Beguš and Zhou (2021a).

$$t = \frac{1}{\|C\|} \sum_{i=1}^{\|C\|} C_i \tag{2}$$

The cABR experiment suggest that peak 2 latency differs significantly between English and Spanish speakers. To test the same acoustic property in deep convolutional layer, we measure peak latency timing between amplitude peaks in output/input and amplitude peaks in the second to last or first convolutional layer (Conv4 or Conv1) in the Generator and Discriminator networks, respectively.

To extract peak timing in each layer, we first generate 20 Generator outputs that are forced to resemble the stimulus (according to Section 3.4) per each model. In three outputs of the first and the second TIMIT replication the output was so unclear that the periodic structure could not be identified, which is why they were removed from the analysis. A total of 74 forced outputs were thus created (2 replications of TIMIT and DIMEx trained models each). For each generated output, we obtain the corresponding representations in the fourth convolutional layer (Conv4) as described in 3.4 by averaging over all feature maps. This yields a time-series data. The generated outputs and time series from the fourth convolutional layers (upsampled) are then annotated for vocalic peaks in Praat (Boersma and Weenink, 2015) and peak timing was extracted with parabolic interpolation.

Peak latency (Δt_n) was calculated as a difference in timing between the peak of the output ($t_{n_{out}}$) and the peak of the fourth convolutional layer ($t_{n_{conv4}}$) in the Generator.

$$\Delta t_n = t_{n_{out}} - t_{n_{conv4}} \tag{3}$$

The burst is annotated as the 0th period and every consecutive period as nth period. The burst is not saliently present in all outputs. A total of 51 bursts (=0th period) were included in the analysis.

To test peak latency in the Discriminator network, we feed the Discriminator the same 74 generated outputs from the Generator forced to resemble the stimulus. The same annotations as for output-Conv4 analysis in the Generator were used to extract peak timing from the forced

[3] Harwath and Glass (2019) propose a visualization technique for the DAVEnet model (Harwath et al., 2020) that involves summation — they operate with L2 norm values of individual filter activations, but they do not operate with the production (decoder) aspect of the networks and operate with spectrograms instead of waveforms. Their visualizations do not offer sufficiently high temporal resolution for comparison with the cABR signal (e.g. for vocalic periods). Their proposal additionally requires a PCA analysis for a comparison of intermediate convolutional layers with linguistically meaningful units. Their model does, however, show, that peak timing in intermediate convolutional layers correspond to segment boundaries (not vocalic peaks) in TIMIT.

generated outputs and the first convolutional layer (Conv1) in the Discriminator network (according to (3)).

4. Results

Visualizations of raw peak latency timing in Figures 3(f,h) and 4(f,h) data suggest that there is a consistent timing difference between the TIMIT-trained models (English) and the DIMEx-trained models (Spanish). The peak latency (Δt_n) is more positive in the Spanish-trained models and more negative in the English-trained models. This observation is consistent in all eight models: in both the Generator and the Discriminator as well as across replications.

4.1. The Generator

To test the significance of the peak latency differences in the Generator, we fit the data to a linear regression model with the PEAK LATENCY timing as the dependent variable and three predictors: LANGUAGE, nTH PERIOD, and REPLICATION with all two-way and three-way interactions.

The LANGUAGE predictor has two levels (English and Spanish) and is treatment-coded with English as the reference level. The nTH PERIOD predictor has 13 levels (for each period and the burst) and is treatment-coded with 1st period as the reference level. Periods above the 12th period are discarded from the analysis due to a small number of attestations (see Figures 3 and 4). REPLICATION is sum-coded with two levels (first and second).

Estimates of the model are given in Table C.5 and Figure 1i. Pairwise comparisons in Table 2 reveal that peak timing does not differ significantly for the burst (0th period) and the first period, but the difference becomes significant for 2nd, 3rd, 4th, and 6th periods (see all estimates in Table 2). If we adjust pairwise comparisons with False Discovery Rate (FDR) adjustment, only differences for the 3rd, 4th, and 6th period are significant. Peak latency differences again become insignificant for the 5th period and periods 7-11. Peak latency differences are significant in individual replications too (See Section Appendix D).

Contrast	nth period	Estimate	SE	df	t.ratio	p.value
English - Spanish	0 (=burst)	0.0000	0.0004	514	0.02	0.981
English - Spanish	1	-0.0006	0.0003	514	-1.91	0.056
English - Spanish	2	-0.0007	0.0003	514	-2.06	0.040
English - Spanish	3	-0.0008	0.0003	514	-2.54	0.011
English - Spanish	4	-0.0010	0.0003	514	-2.83	0.005
English - Spanish	5	-0.0006	0.0004	514	-1.63	0.104
English - Spanish	6	-0.0017	0.0004	514	-4.13	0.000
English - Spanish	7	-0.0005	0.0005	514	-1.09	0.278
English - Spanish	8	-0.0005	0.0008	514	-0.70	0.485
English - Spanish	9	-0.0011	0.0011	514	-1.00	0.318
English - Spanish	10	-0.0021	0.0013	514	-1.63	0.104
English - Spanish	11	-0.0012	0.0014	514	-0.89	0.375

Table 2: Pairwise contrasts in peak timing difference between English and Spanish (pooled across replications) in the Generator network (with *emmeans* package by Lenth 2018). The burst is marked by the 0th period. The 12th period is not estimated due to lack of data.

4.2. The Discriminator

To test the significance of peak latency in the Discriminator network, we perform the same statistical procedure as described in Section 4.1. The peak latency (Δt_n) for the nth period in the Discriminator is calculated as the difference between the peak timing of the input and peak timing of the first convolutional layer (Conv1). The model in Figure 4(i) and Table C.6 includes LANGUAGE, NTH PERIOD, and REPLICATION (coded as in Section 4.1) and all interactions as predictors and the peak latency timing as the dependent variable. The pairwise comparisons in Table 3 show an even bigger difference in peak latency timing between the English- and Spanish-trained Discriminator. Peak latency for the burst (=0th period) does not differ significantly across the two languages. For periods 1–7 (except for 3), the difference is significant (also wit FDR correction); for later periods (and for 3), the difference ceases to be significant again. Peak latency differences are significant in individual replications too (See Section Appendix D).

Contrast	nth period	Estimate	SE	df	t.ratio	p.value
English - Spanish	0	-0.0001	0.0003	514	-0.33	0.745
English - Spanish	1	-0.0007	0.0003	514	-2.64	0.008
English - Spanish	2	-0.0016	0.0003	514	-5.64	0.000
English - Spanish	3	-0.0004	0.0003	514	-1.33	0.183
English - Spanish	4	-0.0019	0.0003	514	-6.77	0.000
English - Spanish	5	-0.0018	0.0003	514	-6.10	0.000
English - Spanish	6	-0.0020	0.0003	514	-5.93	0.000
English - Spanish	7	-0.0012	0.0004	514	-2.87	0.004
English - Spanish	8	-0.0007	0.0006	514	-1.18	0.239
English - Spanish	9	-0.0010	0.0009	514	-1.14	0.253
English - Spanish	10	-0.0018	0.0011	514	-1.66	0.097
English - Spanish	11	-0.0013	0.0011	514	-1.13	0.259

Table 3: Pairwise contrasts in peak timing difference between English and Spanish (pooled across replications) in the Discriminator network (with *emmeans* package by Lenth 2018). The burst is marked by the 0th period. The 12th period is not estimated due to lack of data.

5. Discussion

The results of the computational experiments suggest that peak latency between the forced stimulus and intermediate activations in English-trained and Spanish-trained deep networks differ in similar ways with the peak latency between the stimulus and cABR signal of English and Spanish speakers. We present an interpretable technique that parallels the two modalities. These results raise a question of what properties of deep convolutional networks and the cABR signal cause such parallels.

The main mechanism behind the technique for analyzing acoustic properties in intermediate convolutional layers is a simple averaging of activations of individual feature maps (in equation 2). The second to last convolutional layer (Conv4 in the Generator and Conv1 in the Discriminator) have 64 filters which result in 64 feature maps for each input/output. Individual feature maps offer limited interpretability, but a simple averaged sum over all feature maps after ReLU or Leaky ReLU activation offers highly interpretable time series data (Beguš and Zhou, 2021b,a).

Similar to this proposed computational technique, cABR data represents a summation of neural activity in the brain stem (and potentially also from other non-subcortical sources) (Laumen et al., 2016; Coffey et al., 2019). The basic principle for obtaining the signal between the brainstem and intermediate convolutional layers (as proposed by Beguš and Zhou 2021b) is thus similar.

An analysis of what acoustic information is encoded in the cABR signal and in Conv4/Conv1 also reveals several similarities between the two modalities which provides ground for higher level comparisons conducted in this paper. cABR signals represent several acoustic properties (Kraus and Nicol, 2005; Abrams and Kraus, 2015; BinKhamis et al., 2019): periodicity and the fundamental frequency (F0), lower frequency formants (e.g. F1, perhaps also F2; Krishnan 2002), "acoustic onsets" such as burst, and "frequency transitions" (Abrams and Kraus, 2015). Similarly, it has been shown that the same acoustic properties are encoded in the second to last convolutional layer: periodicity and F0 together with F0 transitions, low frequency formant structure (F1 and to lesser degree F2), burst, and timing of individual segments (Beguš and Zhou, 2021b,a). Figures 1a,b,c,d; 3a,b,c,d; and 4a,b,c,d,e illustrate how the signal from Conv1/4 and cABR encode the same acoustic properties. Higher-level convolutional layers do not encode all these acoustic properties (Beguš and Zhou, 2021b,a). In sum, both the earlier intermediate layers and cABR signal feature encoding of the same acoustic properties (F0, burst, timing, and low frequency formant structure).

Based on these similarities, it is reasonable to assume that both signals represent at least superficially similar computations. Input signal in deep convolutional networks get transformed into spikes in individual feature maps by learned filters. Summing and averaging over these spikes indicates the areas in the layers with most activity and provides an interpretable representation of the input/output. Similarly, the cABR signal summarizes neural activity as a response to the input stimulus.

Peak latency has long been a focus of cABR studies (Kraus and Nicol, 2005; Zhao and Kuhl, 2018; BinKhamis et al., 2019). Zhao and Kuhl (2018) argue that peak 2 latency differs significantly based on language experience, where the two different languages (Spanish and English) have substantially different encoding of a phonetic property which is related to the perception of the sound: voiceless (e.g. [ta]) vs. voiced (e.g. [da]) sounds. This suggests that phonetic features that represent a phonological contrast in language can be encoded early in the auditory pathway—already in the brainstem. Peak latency is an interpretable feature and easy to analyze in deep convolutional networks with the proposed technique that uses summation to identify peak activity in intermediate convolutional layers relative to the input/output. For these reasons, we focus our comparison on peak latency between second to final convolutional layer relative to the input/output and cABR signal in the brain stem relative to the stimulus. Comparison of encoding of other acoustic properties that is made possible by the proposed techniques are left for future work.

The results of the computational experiment suggest that peak amplitude timing of the second to last convolutional layer relative to the speech input/output do not differ significantly for the burst, but do differ significantly for consecutive periods based on what language the models are trained on: English (with long VOT encoding of voicing in stops) and Spanish (without long VOT encoding of voicing stops). The difference is significant both in the Generator (the production principle) as well as in the Discriminator (the perception principle). The peak latency also operates in the same direction across the two replications (in eight models total), which suggests this is not just an idiosyncratic property of individual models. While in the cABR experiment, peak 2 in English speakers precedes peak 2 in Spanish speakers, the timing relationship is the opposite in the computational experiment: the difference in peak amplitude between individual periods in second to last convolutional layer and peak amplitude in the input/output is more positive in the Spanish-trained models.

While the directionality of the difference is the opposite from the brain experiment, the results suggest that a highly interpretable acoustic property — peak latency — that indicates peak activity in the brain stem and in the intermediate convolutional layer relative to the stimulus/input/output based on a common operation, summation/averaging of the signal, is encoded in both in the earlier intermediate convolutional layers and the cABR signal. A more conservative conclusion based on the results is that encoding of speech signal, and more specifically, of peak timing, can differ according to the language exposure (English vs. Spanish) in similar and interpretable ways between the intermediate convolutional layers and the brain stem. Under a less conservative reading of the results, the difference in VOT duration has a similar effect on peak latency in the brain stem and in deep convolutional networks, because peak latency for burst is not significant neither in the brain nor in the intermediate convolutional layers, while subsequent periods show a significant difference in timing in both modalities.

6. Conclusion and future directions

This paper presents a technique for comparing cABR recordings in the brainstem with intermediate convolutional layers in deep neural networks. Both signals are based on summing and averaging of neural activity: either of electrical activity and in the brain stem or of values in individual feature maps in convolutional layers. We show that averaging over feature maps parallels cABR recording in the brain because it summarizes areas in the convolutional layers with highest activity relative to the input/output. cABRs and second to last convolutional layers encode similar acoustic properties. Encoding of phonetic information is tested with cABR experiments on subjects that speak two different languages and with deep neural networks trained on these languages. The results reveal that encoding of phonetic features that result in phonological contrasts differ in similar ways in the brain stem and in intermediate convolutional layers between the two tested languages.

These results provide grounds for comparison of several other acoustic properties using the proposed framework. Both intermediate convolutional layers and cABR signal represent several acoustic properties. World's languages use various acoustic features to encode linguistically meaningful phonological contrasts. Testing these learned representations across different acoustic properties and languages should yield further information on similarities and differences in artificial and biological neural computation on speech data.

Appendix A. Data analysis

All spectrograms are created in Praat Boersma and Weenink (2015) and imported into ggplot2 via a script by Matthew Winn. Peak timing was extracted with a modified script from `https://stackoverflow.com/questions/48138899`. .

Appendix B. GAMMs

Appendix C. Linear models

Appendix D. Individual replications (for Section 4)

Appendix D.1. Generator

In the Generator network, peak latency differs significantly in periods 1, 3, 4, 6 in the first replication and in periods 6 and 10 in the second replication (unadjusted). If adjusted with FDR, period 1 and 6 are significant in the first replication, and period 6 in the second replication.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = English	0.1164	0.0189	6.1428	< 0.0001
LANGUAGE = Spanish	-0.0247	0.0263	-0.9388	0.3486
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Time)	6.1901	6.4494	9.2226	< 0.0001
s(Time):LANGUAGE = Spanish	5.7023	5.9856	2.6980	0.0147
s(Time, Subject)	243.4468	259.0000	205.5821	< 0.0001

Table B.4: Estimates of a generalized additive mixed model (fitted with the *bam()* function in the *mgcv* package by Wood 2011). Amplitude in μV from the EEG-cABR data is the dependent variable. The independent variables include LANGUAGE as a parametric predictor (with two levels, English and Spanish with English treatment-coded as the reference level), smooth for Time, by-Language difference smooth for time, and by-subject random smooths. The model includes correction for autocorrelation.

Appendix D.2. Discriminator

In the Discriminator, periods 1-7 are significant in the first replication and periods 2 and 6 in the second replication (unadjusted). If adjusted with FDR, periods 1-7 are significant in the first replication and period 2 in the second replication.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)=Period1, LanguageEnglish	-0.0010	0.0002	-4.03	0.0001
LanguageSpanish	0.0006	0.0003	1.91	0.0561
Period0	0.0009	0.0004	2.47	0.0137
Period2	0.0003	0.0003	0.89	0.3729
Period3	0.0006	0.0003	1.77	0.0778
Period4	0.0009	0.0003	2.73	0.0065
Period5	0.0010	0.0003	2.76	0.0060
Period6	0.0004	0.0004	1.10	0.2731
Period7	0.0012	0.0004	2.99	0.0030
Period8	0.0007	0.0006	1.17	0.2444
Period9	0.0007	0.0008	0.79	0.4314
Period10	0.0005	0.0008	0.56	0.5787
Period11	0.0011	0.0009	1.16	0.2464
Period12	0.0013	0.0009	1.38	0.1667
Replication11	-0.0008	0.0002	-3.31	0.0010
LanguageSpanish:Period0	-0.0006	0.0005	-1.22	0.2212
LanguageSpanish:Period2	0.0000	0.0005	0.10	0.9188
LanguageSpanish:Period3	0.0002	0.0005	0.43	0.6647
LanguageSpanish:Period4	0.0003	0.0005	0.67	0.5047
LanguageSpanish:Period5	-0.0001	0.0005	-0.13	0.8942
LanguageSpanish:Period6	0.0010	0.0005	1.95	0.0514
LanguageSpanish:Period7	-0.0001	0.0006	-0.18	0.8574
LanguageSpanish:Period8	-0.0001	0.0008	-0.14	0.8897
LanguageSpanish:Period9	0.0004	0.0011	0.38	0.7075
LanguageSpanish:Period10	0.0015	0.0014	1.11	0.2682
LanguageSpanish:Period11	0.0006	0.0014	0.41	0.6831
LanguageSpanish:Period12	0.0002	0.0021	0.11	0.9123
LanguageSpanish:Replication11	0.0009	0.0003	2.82	0.0050
Period0:Replication11	0.0005	0.0004	1.37	0.1704
Period2:Replication11	0.0008	0.0003	2.27	0.0233
Period3:Replication11	0.0006	0.0003	1.86	0.0631
Period4:Replication11	0.0007	0.0003	2.02	0.0439
Period5:Replication11	0.0008	0.0003	2.30	0.0218
Period6:Replication11	0.0007	0.0004	1.98	0.0478
Period7:Replication11	0.0006	0.0004	1.37	0.1725
Period8:Replication11	0.0009	0.0006	1.43	0.1531
Period9:Replication11	0.0010	0.0008	1.14	0.2548
Period10:Replication11	0.0013	0.0008	1.58	0.1142
Period11:Replication11	0.0007	0.0009	0.78	0.4347
Period12:Replication11	0.0009	0.0009	1.01	0.3125
LanguageSpanish:Period0:Replication11	-0.0005	0.0005	-0.87	0.3839
LanguageSpanish:Period2:Replication11	-0.0007	0.0005	-1.54	0.1231
LanguageSpanish:Period3:Replication11	-0.0007	0.0005	-1.57	0.1174
LanguageSpanish:Period4:Replication11	-0.0008	0.0005	-1.72	0.0866
LanguageSpanish:Period5:Replication11	-0.0013	0.0005	-2.59	0.0099
LanguageSpanish:Period6:Replication11	-0.0010	0.0005	-1.91	0.0565
LanguageSpanish:Period7:Replication11	-0.0003	0.0006	-0.56	0.5788
LanguageSpanish:Period8:Replication11	-0.0006	0.0008	-0.77	0.4414
LanguageSpanish:Period9:Replication11	-0.0016	0.0011	-1.48	0.1395
LanguageSpanish:Period10:Replication11	-0.0028	0.0014	-2.09	0.0371
LanguageSpanish:Period11:Replication11	-0.0005	0.0014	-0.34	0.7316

Table C.5: Estimates of the linear model described in Section 4.1 with three predictors (LANGUAGE, NTH PERIOD, and REPLICATION with all two-way and three-way interactions) for the Generator network.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)=Period1, LanguageEnglish	-0.0001	0.0002	-0.60	0.5485
LanguageSpanish	0.0007	0.0003	2.64	0.0084
Period0	0.0004	0.0003	1.25	0.2133
Period2	-0.0005	0.0003	-1.63	0.1027
Period3	-0.0001	0.0003	-0.36	0.7221
Period4	-0.0001	0.0003	-0.23	0.8166
Period5	0.0000	0.0003	0.12	0.9066
Period6	-0.0003	0.0003	-1.00	0.3201
Period7	-0.0002	0.0003	-0.53	0.5963
Period8	-0.0001	0.0005	-0.14	0.8912
Period9	-0.0000	0.0007	-0.06	0.9492
Period10	-0.0000	0.0007	-0.07	0.9476
Period11	0.0000	0.0008	0.05	0.9570
Period12	-0.0002	0.0008	-0.26	0.7937
Replication11	-0.0003	0.0002	-1.50	0.1334
LanguageSpanish:Period0	-0.0006	0.0004	-1.41	0.1583
LanguageSpanish:Period2	0.0008	0.0004	2.12	0.0348
LanguageSpanish:Period3	-0.0004	0.0004	-0.93	0.3503
LanguageSpanish:Period4	0.0012	0.0004	2.96	0.0032
LanguageSpanish:Period5	0.0011	0.0004	2.61	0.0093
LanguageSpanish:Period6	0.0013	0.0004	2.88	0.0042
LanguageSpanish:Period7	0.0004	0.0005	0.89	0.3757
LanguageSpanish:Period8	0.0000	0.0007	0.00	0.9963
LanguageSpanish:Period9	0.0003	0.0009	0.29	0.7703
LanguageSpanish:Period10	0.0011	0.0011	0.96	0.3368
LanguageSpanish:Period11	0.0006	0.0012	0.47	0.6382
LanguageSpanish:Period12	0.0016	0.0018	0.93	0.3513
LanguageSpanish:Replication11	0.0004	0.0003	1.29	0.1973
Period0:Replication11	0.0002	0.0003	0.54	0.5907
Period2:Replication11	0.0002	0.0003	0.57	0.5688
Period3:Replication11	0.0001	0.0003	0.18	0.8558
Period4:Replication11	-0.0000	0.0003	-0.01	0.9882
Period5:Replication11	-0.0000	0.0003	-0.05	0.9619
Period6:Replication11	0.0002	0.0003	0.55	0.5799
Period7:Replication11	-0.0001	0.0003	-0.15	0.8773
Period8:Replication11	0.0001	0.0005	0.19	0.8491
Period9:Replication11	0.0001	0.0007	0.07	0.9419
Period10:Replication11	0.0000	0.0007	0.03	0.9738
Period11:Replication11	-0.0001	0.0008	-0.09	0.9269
Period12:Replication11	0.0001	0.0008	0.18	0.8539
LanguageSpanish:Period0:Replication11	-0.0001	0.0004	-0.18	0.8596
LanguageSpanish:Period2:Replication11	0.0001	0.0004	0.16	0.8735
LanguageSpanish:Period3:Replication11	0.0004	0.0004	1.14	0.2558
LanguageSpanish:Period4:Replication11	0.0010	0.0004	2.45	0.0146
LanguageSpanish:Period5:Replication11	0.0009	0.0004	2.12	0.0343
LanguageSpanish:Period6:Replication11	0.0006	0.0004	1.41	0.1597
LanguageSpanish:Period7:Replication11	0.0008	0.0005	1.56	0.1183
LanguageSpanish:Period8:Replication11	0.0003	0.0007	0.45	0.6551
LanguageSpanish:Period9:Replication11	-0.0008	0.0009	-0.85	0.3963
LanguageSpanish:Period10:Replication11	-0.0009	0.0011	-0.83	0.4056
LanguageSpanish:Period11:Replication11	0.0008	0.0012	0.68	0.4941

Table C.6: Estimates of the linear model described in Section 4.2 with three predictors (language, nth period, and replication with all two-way and three-way interactions) for the Discriminator network.

Period	Replication	contrast	estimate	SE	df	t.ratio	p.value
0	1	English - Spanish	-0.0005	0.0006	514	-0.797	0.6516
1	1	English - Spanish	-0.0016	0.0005	514	-3.419	0.0177
2	1	English - Spanish	-0.0009	0.0005	514	-1.945	0.1944
3	1	English - Spanish	-0.0010	0.0005	514	-2.263	0.1043
4	1	English - Spanish	-0.0011	0.0005	514	-2.339	0.1025
5	1	English - Spanish	-0.0003	0.0005	514	-0.550	0.7569
6	1	English - Spanish	-0.0016	0.0005	514	-3.049	0.0314
7	1	English - Spanish	-0.0011	0.0006	514	-1.832	0.2197
8	1	English - Spanish	-0.0008	0.0012	514	-0.710	0.6907
9	1	English - Spanish	-0.0004	0.0017	514	-0.212	0.9407
10	1	English - Spanish	-0.0002	0.0021	514	-0.119	0.9543
11	1	English - Spanish	-0.0017	0.0021	514	-0.809	0.6516
12	1	English - Spanish	-0.0018	0.0021	514	-0.877	0.6516
0	2	English - Spanish	0.0005	0.0006	514	0.850	0.6516
1	2	English - Spanish	0.0003	0.0005	514	0.625	0.7285
2	2	English - Spanish	-0.0005	0.0005	514	-0.988	0.6474
3	2	English - Spanish	-0.0006	0.0005	514	-1.345	0.3881
4	2	English - Spanish	-0.0008	0.0005	514	-1.684	0.2413
5	2	English - Spanish	-0.0009	0.0005	514	-1.691	0.2413
6	2	English - Spanish	-0.0017	0.0006	514	-2.819	0.0434
7	2	English - Spanish	0.0001	0.0008	514	0.104	0.9543
8	2	English - Spanish	-0.0002	0.0009	514	-0.232	0.9407
9	2	English - Spanish	-0.0017	0.0013	514	-1.384	0.3881
10	2	English - Spanish	-0.0040	0.0016	514	-2.476	0.0885
11	2	English - Spanish	-0.0008	0.0018	514	-0.423	0.8328
12	2	English - Spanish					NA

P-value adjustment: FDR method for 26 tests

Table D.7: Pairwise contrasts in peak timing difference between English and Spanish across replications in the Generator network with FDR adjustment (with *emmeans* package by Lenth 2018). The burst is marked by the 0th period. The 12th period is not estimated due to lack of data.

Period	Replication	contrast	estimate	SE	df	t.ratio	p.value
0	1	English - Spanish	-0.0004	0.0005	514	-0.801	0.5798
1	1	English - Spanish	-0.0011	0.0004	514	-2.843	0.0151
2	1	English - Spanish	-0.0020	0.0004	514	-5.168	<.0001
3	1	English - Spanish	-0.0012	0.0004	514	-3.050	0.0104
4	1	English - Spanish	-0.0032	0.0004	514	-8.410	<.0001
5	1	English - Spanish	-0.0030	0.0004	514	-7.716	<.0001
6	1	English - Spanish	-0.0030	0.0004	514	-6.769	<.0001
7	1	English - Spanish	-0.0023	0.0005	514	-4.433	0.0001
8	1	English - Spanish	-0.0014	0.0010	514	-1.438	0.2990
9	1	English - Spanish	-0.0006	0.0014	514	-0.414	0.8411
10	1	English - Spanish	-0.0012	0.0017	514	-0.721	0.6130
11	1	English - Spanish	-0.0024	0.0017	514	-1.422	0.2990
12	1	English - Spanish	-0.0027	0.0017	514	-1.592	0.2649
0	2	English - Spanish	0.0002	0.0005	514	0.354	0.8554
1	2	English - Spanish	-0.0004	0.0004	514	-0.937	0.5044
2	2	English - Spanish	-0.0011	0.0004	514	-2.854	0.0151
3	2	English - Spanish	0.0004	0.0004	514	1.099	0.4163
4	2	English - Spanish	-0.0006	0.0004	514	-1.404	0.2990
5	2	English - Spanish	-0.0006	0.0004	514	-1.306	0.3122
6	2	English - Spanish	-0.0010	0.0005	514	-2.001	0.1327
7	2	English - Spanish	-0.0000	0.0006	514	-0.065	0.9859
8	2	English - Spanish	-0.0001	0.0008	514	-0.093	0.9859
9	2	English - Spanish	-0.0014	0.0011	514	-1.354	0.3054
10	2	English - Spanish	-0.0024	0.0014	514	-1.768	0.2017
11	2	English - Spanish	-0.0001	0.0015	514	-0.084	0.9859
12	2	English - Spanish					NA

P-value adjustment: FDR method for 26 tests

Table D.8: Pairwise contrasts in peak timing difference between English and Spanish across replications in the Discriminator network with FDR adjustment (with *emmeans* package by Lenth 2018). The burst is marked by the 0th period. The 12th period is not estimated due to lack of data.

References

Abrams, D.A., Kraus, N., 2015. Auditory pathway representations of speech sounds in humans, in: Handbook of Clinical Audiology. Wolters Kluwer Health. chapter 28, pp. 527–544.

Beguš, G., 2019. Post-nasal devoicing and the blurring process. Journal of Linguistics 55, 689–753. doi:10.1017/S002222671800049X.

Beguš, G., 2020a. Estimating historical probabilities of natural and unnatural processes. Phonology 37, 515–549. doi:10.1017/S0952675720000263.

Beguš, G., 2020b. Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. Frontiers in Artificial Intelligence 3, 44. URL: https://www.frontiersin.org/article/10.3389/frai.2020.00044, doi:10.3389/frai.2020.00044.

Beguš, G., 2021a. Ciwgan and fiwgan: Encoding information in acoustic data to model lexical learning with generative adversarial networks. Neural Networks 139, 305–325. URL: https://www.sciencedirect.com/science/article/pii/S0893608021001052, doi:https://doi.org/10.1016/j.neunet.2021.03.017.

Beguš, G., 2021b. Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication. Transactions of the Association for Computational Linguistics 9, 1180–1196. URL: https://doi.org/10.1162/tacl_a_00421, doi:10.1162/tacl_a_00421, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00421/1971814/tacl_a_00421.pd

Beguš, G., 2021c. Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. Computer Speech & Language , 101244URL: https://www.sciencedirect.com/science/article/pii/S0885230821000516, doi:https://doi.org/10.1016/j.csl.2021.101244.

Beguš, G., Zhou, A., 2021a. Interpreting intermediate convolutional layers in unsupervised acoustic word classification. ArXiv abs/2110.02375.

Beguš, G., Zhou, A., 2021b. Interpreting intermediate convolutional layers of cnns trained on raw speech. CoRR abs/2104.09489. URL: https://arxiv.org/abs/2104.09489, arXiv:2104.09489.

Bengio, Y., Lee, D.H., Bornschein, J., Mesnard, T., Lin, Z., 2016. Towards biologically plausible deep learning. arXiv:1502.04156.

Bidelman, G.M., Gandour, J.T., Krishnan, A., 2011. Cross-domain Effects of Music and Language Experience on the Representation of Pitch in the Human Auditory Brainstem. Journal of Cognitive Neuroscience 23, 425–434. URL: https://doi.org/10.1162/jocn.2009.21362, doi:10.1162/jocn.2009.21362, arXiv:https://direct.mit.edu/jocn/article-pdf/23/2/425/1940751/jocn.2009.21362.pdf.

BinKhamis, G., Léger, A., Bell, S.L., Prendergast, G., O'Driscoll, M., Kluk, K., 2019. Speech auditory brainstem responses: Effects of background, stimulus duration, consonant–vowel, and number of epochs. Ear and Hearing 40. URL: https://journals.lww.com/ear-hearing/Fulltext/2019/05000/Speech_Auditory_Brainstem_Responses__Effects_of.21.aspx.

Blevins, J., 2013. Evolutionary Phonology: A holistic approach to sound change typology, in: Honeybone, P., Salmons, J. (Eds.), Handbook of Historical Phonology. Oxford University Press, Oxford.

Boersma, P., Weenink, D., 2015. Praat: doing phonetics by computer [computer program]. version 5.4.06. Retrieved 21 February 2015 from http://www.praat.org/.

Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J., 2014. Deep neural networks rival the representation of primate it cortex for core visual object recognition. PLOS Computational Biology 10, 1–18. URL: https://doi.org/10.1371/journal.pcbi.1003963, doi:10.1371/journal.pcbi.1003963.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports 6, 27755. URL: https://doi.org/10.1038/srep27755, doi:10.1038/srep27755.

Coffey, E.B.J., Nicol, T., White-Schwoch, T., Chandrasekaran, B., Krizman, J., Skoe, E., Zatorre, R.J., Kraus, N., 2019. Evolving perspectives on the sources of the frequency-following response. Nature Communications 10, 5036. URL: https://doi.org/10.1038/s41467-019-13003-w, doi:10.1038/s41467-019-13003-w.

Culbertson, J., Kirby, S., 2016. Simplicity and specificity in language: Domain-general biases have domain-specific effects. Frontiers in Psychology 6, 1964. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2015.01964, doi:10.3389/fpsyg.2015.01964.

Donahue, C., Balsubramani, A., McAuley, J.J., Lipton, Z.C., 2017. Semantically decomposing the latent spaces of generative adversarial networks. CoRR abs/1705.07904. URL: http://arxiv.org/abs/1705.07904, arXiv:1705.07904.

Donahue, C., McAuley, J.J., Puckette, M.S., 2019. Adversarial audio synthesis, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net. URL: https://openreview.net/forum?id=ByMVTsR5KQ.

Donhauser, P.W., Baillet, S., 2020. Two distinct neural timescales for predictive speech processing. Neuron 105, 385–393.e9. URL: https://www.sciencedirect.com/science/article/pii/S0896627319308931, doi:https://doi.org/10.1016/j.neuron.2019.10.019.

Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2017. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage 152, 184–194. URL: https://www.sciencedirect.com/science/article/pii/S1053811916305481, doi:https://doi.org/10.1016/j.neuroimage.2016.10.001.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36, 193–202. URL: https://doi.org/10.1007/BF00344251, doi:10.1007/BF00344251.

Garofolo, J.S., Lamel, L., M Fisher, W., Fiscus, J., S. Pallett, D., L. Dahlgren, N., Zue, V., 1993. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium .

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Greene, M.R., Hansen, B.C., 2018. Shared spatiotemporal category representations in biological and artificial deep neural networks. PLOS Computational Biology 14, 1–17. URL: https://doi.org/10.1371/journal.pcbi.1006327, doi:10.1371/journal.pcbi.1006327.

Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience 35, 10005–10014. URL: https://www.jneurosci.org/content/35/27/10005, doi:10.1523/JNEUROSCI.5023-14.2015, arXiv:https://www.jneurosci.org/content/35/27/10005.full.pdf.

Guest, O., Martin, A.E., 2021. On logical inference over brains, behaviour, and artificial neural networks. URL: psyarxiv.com/tbmcg, doi:10.31234/osf.io/tbmcg.

Harwath, D., Glass, J., 2019. Towards visually grounded sub-word speech unit discovery, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3017–3021. doi:10.1109/ICASSP.2019.8682666.

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J., 2020. Jointly discovering visual objects and spoken words from raw sensory input. International Journal of Computer Vision 128, 620–641. URL: https://doi.org/10.1007/s11263-019-01205-0, doi:10.1007/s11263-019-01205-0.

Huang, N., Slaney, M., Elhilali, M., 2018. Connecting deep neural networks to physical, perceptual, and electrophysiological auditory signals. Frontiers in Neuroscience 12, 532. URL: https://www.frontiersin.org/article/10.3389/fnins.2018.00532, doi:10.3389/fnins.2018.00532.

Jain, S., Huth, A., 2018. Incorporating context into language encoding models for fmri, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Jat, S., Tang, H., Talukdar, P., Mitchell, T., 2019. Relating simple sentence representations in deep neural networks and the brain, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy. pp. 5137–5154. URL: https://aclanthology.org/P19-1507, doi:10.18653/v1/P19-1507.

Kell, A.J., McDermott, J.H., 2019. Deep neural network models of sensory systems: windows onto the role of task constraints. Current Opinion in Neurobiology 55, 121–132. URL: https://www.sciencedirect.com/science/article/pii/S0959438818302034, doi:https://doi.org/10.1016/j.conb.2019.02.003. machine Learning, Big Data, and Neuroscience.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H., 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, 630–644.e16. URL: https://doi.org/10.1016/j.neuron.2018.03.044, doi:10.1016/j.neuron.2018.03.044.

Keyes, A., Bayat, N., Khazaie, V.R., Mohsenzadeh, Y., 2020. Latent Vector Recovery of Audio GANs. arXiv URL: https://arxiv.org/abs/2010.08534v1, arXiv:2010.08534.

Khatami, F., Escabí, M.A., 2020. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. PLOS Computational Biology 16, 1–27. URL: https://doi.org/10.1371/journal.pcbi.1007558, doi:10.1371/journal.pcbi.1007558.

Kiparsky, P., 2006. Amphichronic program vs. Evolutionary Phonology. Theoretical Linguistics 32, 217–236.

Kiparsky, P., 2008. Universals constrain change, change results in typological generalizations, in: Good, J. (Ed.), Linguistic universals and language change. Oxford University Press, Oxford, pp. 23–53.

Koumura, T., Terashima, H., Furukawa, S., 2019. Cascaded tuning to amplitude modulation for natural sound recognition. Journal of Neuroscience 39, 5517–5533. URL: https://www.jneurosci.org/content/39/28/5517, doi:10.1523/JNEUROSCI.2914-18.2019, arXiv:https://www.jneurosci.org/content/39/28/5517.full.pdf.

Kraus, N., Nicol, T., 2005. Brainstem origins for cortical 'what' and 'where' pathways in the auditory system. Trends in Neurosciences 28, 176–181. URL: https://www.sciencedirect.com/science/article/pii/S0166223605000470, doi:https://doi.org/10.1016/j.tins.2005.02.003.

Kriegeskorte, N., Douglas, P.K., 2019. Interpreting encoding and decoding models. Current Opinion in Neurobiology 55, 167–179. URL: https://www.sciencedirect.com/science/article/pii/S0959438818301004, doi:https://doi.org/10.1016/j.conb.2019.04.002. machine Learning, Big Data, and Neuroscience.

Krishnan, A., 2002. Human frequency-following responses: representation of steady-state synthetic vowels. Hearing Research 166, 192–201. URL: https://www.sciencedirect.com/science/article/pii/S0378595502003271, doi:https://doi.org/10.1016/S0378-5955(02)00327-1.

Laumen, G., Ferber, A.T., Klump, G.M., Tollin, D.J., 2016. The physiological basis and clinical use of the binaural interaction component of the auditory brainstem response. Ear and Hearing 37. URL: https://journals.lww.com/ear-hearing/Fulltext/2016/09000/The_Physiological_Basis_and_Clinical_Use_of_the.14.aspx.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Computation 1, 541–551. doi:10.1162/neco.1989.1.4.541.

Lenth, R., 2018. emmeans: Estimated Marginal Means, aka Least-Squares Means. URL: https://CRAN.R-project.org/package=emmeans. r package version 1.3.0.

Lindsay, G.W., 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. Journal of Cognitive Neuroscience 33, 2017–2031. URL: https://doi.org/10.1162/jocn_a_01544, doi:10.1162/jocn_a_01544, arXiv:https://direct.mit.edu/jocn/article-pdf/33/10/2017/1962083/jocn_a_01544.pdf.

Lipton, Z.C., Tripathi, S., 2017. Precise Recovery of Latent Vectors from Generative Adversarial Networks. arXiv URL: https://arxiv.org/abs/1702.04782v2, arXiv:1702.04782.

Marblestone, A.H., Wayne, G., Kording, K.P., 2016. Toward an integration of deep learning and neuroscience. Frontiers in Computational Neuroscience 10, 94. URL: https://www.frontiersin.org/article/10.3389/fncom.2016.00094, doi:10.3389/fncom.2016.00094.

Millet, J., King, J.R., 2021. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. arXiv:2103.01032.

Norman-Haignere, S.V., McDermott, J.H., 2018. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLOS Biology 16, 1–46. URL: https://doi.org/10.1371/journal.pbio.2005127, doi:10.1371/journal.pbio.2005127.

Piantadosi, S.T., Fedorenko, E., 2017. Infinitely productive language can arise from chance under communicative pressure. Journal of Language Evolution 2, 141–147. URL: https://doi.org/10.1093/jole/lzw013, doi:10.1093/jole/lzw013, arXiv:https://academic.oup.com/jole/article-pdf/2/2/141/29093162/lzw013.pdf.

Pineda, L.A., Pineda, L.V., Cuétara, J., Castellanos, H., López, I., 2004. DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish, in: Advances in Artificial Intelligence – IBERAMIA 2004. Springer, Berlin, Germany, pp. 974–983. doi:10.1007/978-3-540-30498-2_97.

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M.R., Wennekers, T., 2021. Biological constraints on neural network models of cognitive function. Nature Reviews Neuroscience 22, 488–502. URL: https://doi.org/10.1038/s41583-021-00473-5, doi:10.1038/s41583-021-00473-5.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 .

Saddler, M.R., Gonzalez, R., McDermott, J.H., 2021. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. Nature Communications 12, 7278. URL: https://doi.org/10.1038/s41467-021-27366-6, doi:10.1038/s41467-021-27366-6.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J., Fedorenko, E., 2020. Artificial neural networks accurately predict language processing in the brain. bioRxiv URL: https://www.biorxiv.org/content/early/2020/06/27/2020.06.26.174482, doi:10.1101/2020.06.26.174482, arXiv:https://www.biorxiv.org/content/early/2020/06/27/2020.06.26.174482.full.pdf.

Skoe, E., Kraus, N., 2010. Auditory brain stem response to complex sounds: A tutorial. Ear and Hearing 31. URL: https://journals.lww.com/ear-hearing/Fulltext/2010/06000/Auditory_Brain_Stem_Response_to_Complex_Sounds__A.2.aspx.

Smith, S.S., Sollini, J., Akeroyd, M.A., 2021a. Inferring the neural basis of binaural detection using deep learning. bioRxiv URL: https://www.biorxiv.org/content/early/2021/01/05/2021.01.05.425246, doi:10.1101/2021.01.05.425246, arXiv:https://www.biorxiv.org/content/early/2021/01/05/2021.01.05.425246.full.pdf.

Smith, S.S., Sollini, J., Akeroyd, M.A., 2021b. Inferring the neural basis of binaural phenomena with a modified autoencoder. bioRxiv URL: https://www.biorxiv.org/content/early/2021/05/13/2021.01.05.425246, doi:10.1101/2021.01.05.425246, arXiv:https://www.biorxiv.org/content/early/2021/05/13/2021.01.05.425246.full.pdf.

Storrs, K.R., Kriegeskorte, N., 2019. Deep learning for cognitive neuroscience. arXiv:1903.01458.

la Tour, T.D., Lu, M., Eickenberg, M., Gallant, J.L., 2021. A finer mapping of convolutional neural network layers to the visual cortex, in: SVRHM 2021 Workshop @ NeurIPS. URL: https://openreview.net/forum?id=EcoKpq43Ul8.

Vihman, M., 2015. Perception and production in phonological development, in: The Handbook of Language Emergence. John Wiley & Sons, Ltd. chapter 20, pp. 437–457. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118346136.ch20, doi:https://doi.org/10.1002/9781118346136.ch20, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118346136.ch20.

Whittington, J.C., Bogacz, R., 2019. Theories of error back-propagation in the brain. Trends in Cognitive Sciences 23, 235–250. URL: https://www.sciencedirect.com/science/article/pii/S1364661319300129, doi:https://doi.org/10.1016/j.tics.2018.12.005.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73, 3–36.

Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience 19, 356–365. URL: https://doi.org/10.1038/nn.4244, doi:10.1038/nn.4244.

Zhao, T.C., Kuhl, P.K., 2018. Linguistic effect on speech perception observed at the brainstem. Proceedings of the National Academy of Sciences 115, 8716–8721. URL: https://www.pnas.org/content/115/35/8716, doi:10.1073/pnas.1800186115, arXiv:https://www.pnas.org/content/115/35/8716.full.pdf.

Zhao, T.C., Masapollo, M., Polka, L., Ménard, L., Kuhl, P.K., 2019. Effects of formant proximity and stimulus prototypicality on the neural discrimination of vowels: Evidence from the auditory frequency-following response. Brain and Language 194, 77–83. URL: https://www.sciencedirect.com/science/article/pii/S0093934X1930032X, doi:https://doi.org/10.1016/j.bandl.2019.05.002.