# REPRESENTING SPEECH THROUGH AUTOREGRESSIVE PREDICTION OF COCHLEAR TOKENS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We introduce a biologically-inspired model for encoding speech through an autoregressive prediction objective on input representations modeled after the human cochlea. Our modeling framework is inspired by the human auditory processing hierarchy. The first stage of our framework transforms the raw audio waveform into a time-frequency representation based on the human cochlea, with an intermediate stage that bottlenecks the audio into discrete units which we denote as *cochlear tokens*. The second stage of our framework learns a simple, yet powerful, autoregressive sequence model over the cochlear tokens. We demonstrate that our model learns meaningful representations of phonemes and word identities, and state-of-the-art representations of lexical semantics. In addition, our model shows competitive performance on a diverse set of downstream speech tasks from the SUPERB benchmark. Complementing our model's strong representational capabilities, we demonstrate its abilities to generate continuations of audio at various temporal scales, which can be visualized in a spectrogram space to provide insights into the model's predictions. Our model provides a novel framework for speech representation learning, aiming to advance the development of more human-like models that flexibly and efficiently handles a range of speech-based tasks.

## 1 INTRODUCTION

Humans possess a remarkable ability to perform a wide range of distinct tasks from speech–ranging from recognizing words in noisy environments and separating multiple speakers to identifying who is speaking and interpreting the emotional tone of their voice. These highly distinct processes are carried out by the human ear and networks of biological neurons. However, building artificial neural network models that replicate the human ability to flexibly and efficiently understand and interact with the world through speech in highly diverse ways remains a significant challenge (1; 2; 3). **To bridge this gap, we introduce CochStream, a biologically-inspired model that learns versatile speech representations through a simple and scalable autoregressive prediction objective on a time-frequency representation inspired by the human cochlea (4; 5; 6; 7).**

### 1.1 SPEECH REPRESENTATION LEARNING

Speech representation models, also known as audio encoders, broadly take an audio signal as input and embed it in a series of discrete tokens or continuous embeddings for use in a variety of downstream audio tasks (3). Several approaches exist for speech representation learning. One popular approach uses neural audio codecs, which learn compressed audio representations by preserving a minimal amount of information needed to reconstruct an audio signal. This compression allows codec models to naturally recover the original signal from the learned codes (8; 9; 10; 11; 12; 13; 14; 15; 16). Many audio codec models are specific to speech (14; 17; 13; 15) while others are more general and include other audio categories such as music, environmental sounds, etc. 10; 8. Models such as SoundStream (8) achieve impressive bit rates (as low as 3kbps) while maintaining high reconstruction quality. These audio codes can then be used as representation for downstream audio tasks (18; 19; 2; 15). However, although these models retain high-fidelity information about the acoustic details due to the reconstruction objective, learning the appropriate acoustic invariances continues to be a challenge (2). Further, high-fidelity signal reconstruction is not a biologically plausible objective – humans

show robust invariance to many low-level features which for instance allows us to understand diverse speakers (20; 21).

A second popular approach for speech representation learning is prediction-based modeling. These models predict features derived either from the raw waveform (22; 23; 24) or from a time-frequency representation of the audio (25; 26; 27; 28). Broadly, these prediction-based speech models fall into two categories: those that predict future frames with an autoregressive objective (25; 26; 27; 28) or those that predict masked frames from surrounding frames (24; 29; 22; 30) (analogous to the causal and bi-directional prediction approaches in language modeling). The resulting representations are then utilized in various downstream audio tasks, for instance language modeling (30; 22; 31; 32; 33; 34). One of the most widely used predictive models is HuBERT (22) which adapts the bi-directional BERT (35) objective for speech representation learning via waveform features.

A third common approach is contrastive learning, where frames from different audio samples are pushed together or pulled apart in the embedding space based on a specified objective. One popular model within contrastive models is wav2vec2 (36) which contrasts masked-out audio segments from distractors in combination with an auxiliary objective. Other models that utilize contrastive objectives include CPC (37), SCPC (38), and COLA (39). Broadly, this approach can yield powerful representations but requires a heuristic to determine positive and negative samples, implicitly enforcing which aspects of the audio signal are retained. Moreover, contrastive objectives often rely on directly contrasting embeddings across hundreds or thousands of diverse samples simultaneously, which, arguably, is not a biologically plausible operation.

Although these three speech representation learning strategies are distinct, their objectives can be combined and augmented with additional heuristics. For instance, a state-of-the-art model, WavLM (23), combines the HuBERT bi-directional prediction objective (22) with a noisy input transformation to obtain strong numbers on the speech representation SUPERB benchmark (1). However, as with most ensemble models, these performance gains come at the cost of additional hand-crafted complexity.

### 1.2 Our Approach: A Two-Stage Framework for Autoregressive Prediction on Biologically-Inspired Input Representations

In contrast to these past approaches, our proposed framework does not rely on signal-reconstruction objectives (like neural codec models), non-causal prediction objectives (like bi-directional prediction models), or intra-batch contrasting of samples (like many contrastive models). Instead, our framework takes inspiration from the human auditory processing stream and operates in two stages:

The first stage of our framework transforms the raw audio waveform into a time-frequency representation based on the human cochlea (**WavCoch**; Figure 1A-C). Note that this approach in some way resembles neural audio codecs, but instead of reconstructing the *same* signal, we predict another audio representation–one known to be computed within the human auditory processing hierarchy—the time-frequency cochleagram (4; 5; 6; 7). We probe the representations in an intermediate bottleneck stage of WavCoch which effectively discretizes the audio representations (Section 2.1.1). We refer to these discrete, intermediary representations as **cochlear tokens** (Figure 1B), and broadly term this strategy of training a computational model to transition between representational states while extracting intermediary representations as *"Transformation Imitation"* (see Section 4). We hypothesize that the cochlear tokens learned through this transformation will develop inductive biases similar to those of the human cochlea, benefiting the model's ability to process diverse speech patterns efficiently (e.g., 40).

The cochlear tokens serve as input to the second stage of our framework, **CochStream**, which is a simple autoregressive sequence model, trained to predict the upcoming cochlear token (Figure 1D) (Section 2.1.2). Since the cochlear tokens were derived from a waveform-to-cochleagram transformation, these predicted cochlear tokens can naturally be decoded into the cochleagram representation for inspection and interpretability (Figure 1E). Compared to existing autoregressive prediction models, we do not jointly learn to embed and predict future frames, but rather utilize our biologically-inspired WavCoch quantization model to identify tokens, and subsequently learn to model a distribution over this fixed token vocabulary via autoregressive prediction.

Hence, we formulate speech representation learning through a simple, yet powerful, autoregressive prediction objective on biologically-realistic input representations, cochlear tokens. We demonstrate

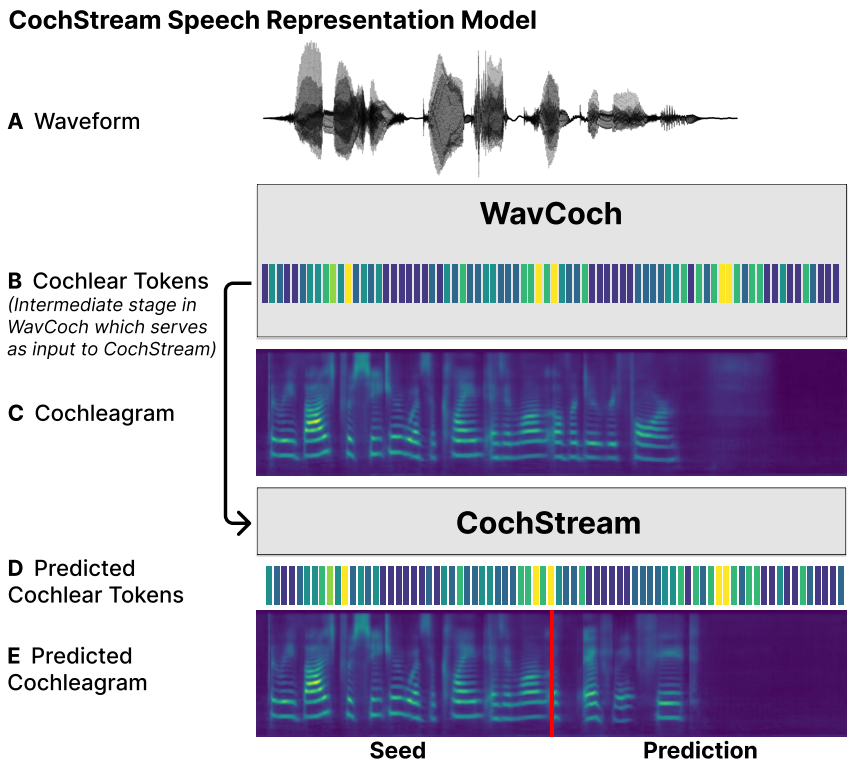


Figure 1: **Schematic of CochStream speech representation model**. Description of the steps associated with the model can be found in the Introduction (1.2) and Methods (2.1).

.

that our framework leads to the emergence of representations from which phonemes, word forms, and word meanings (lexical semantics) can be decoded at competitive levels, when compared to several comparison models (Section 3.1 and 3.2), with our approach obtaining state-of-the-art performance on lexical semantics (Section 3.1). Further, we show that our learned representations serve as a a powerful backbone for various downstream speech tasks, as evaluated on the SUPERB benchmark (1) (Section 3.3). Finally, unlike the comparison models, we demonstrate CochStream's capability to generate continuations of audio when prompted, which can be visualized in a cochleagram time-frequency space to provide insights into the model's predictions (Section 3).

**So, why should we care about our novel framework? From a machine learning perspective**, we introduce cochlear tokens which fit within the context window of a standard Transformer, effectively leveraging the power of autoregressive modeling (41). By casting speech representation learning as a generic sequence prediction task, our model matches the performance of comparably sized models. Yet, with the scalable nature of autoregressive models (demonstrated in Section 3.1 and 3.3) and availability of large amounts of unlabeled speech data, our approach is well-positioned to outperform existing methods at larger scales. **Within cognitive science and neuroscience**, a growing body of research uses artificial models to investigate human behavior and neural processes, under the premise that models operating on realistic input signals and/or solving biologically relevant tasks may develop representations or computations resembling those of biological systems (42; 43). Recent work has investigated whether these models mimic human behavior (44; 40; 45; 46; 47) and neural activity (48; 49; 50; 51; 52). However, current artificial models fall short in matching key aspects of biological systems (53; 54; 7). Hence, developing new classes of models that incorporate biological heuristics, such as a cochlear-inspired input representations, is crucial for advancing accurate computational accounts of human cognition and brain function.

# 2 METHODS

## 2.1 COCHSTREAM MODEL

The CochStream model (Section 2.1.2) is trained to autoregressively predict a sequence of cochlear tokens produced by the WavCoch encoder (Section 2.1.1).

### 2.1.1 INPUT REPRESENTATIONS: WAVCOCH

We propose a method for efficiently tokenizing audio through a waveform-to-cochleagram transformation model. This model, **WavCoch**, loosely mimics the function of the human cochlea, transforming auditory inputs into a time-frequency representation (4; 5; 6; 7).

WavCoch is a vector-quantized encoder which takes as input 5s clips of mono audio waveforms sampled at 16 kHz and is trained to predict a time-frequency representation of the corresponding audio clip. The target time-frequency representation is a human-inspired cochleagram representation (55; 56; 7), consisting of 221 frequency bins and 988 temporal steps (7). The purpose of the WavCoch model is to extract discrete tokens from continuous audio signal to serve as the input to CochStream. The full model diagram is illustrated in Appendix Figure 1 and the steps are described below.

First, the raw waveform (shape: [1,80000] for 5s of mono audio sampled at 16kHz) undergoes the Fourier Transform by computing Twiddle Factors (57). These factors represent complex sinusoidal components that decompose the signal into its frequency spectrum. The Twiddle Factors are applied to the audio signal through a 1D convolution (window size 1,001 and hop length 80 samples) which transforms the signal into the time-frequency domain. Second, each 5 ms temporal step of this time-frequency representation is fed into two fully-connected (FC) layers with ReLU nonlinearities (with 512 hidden units each). Third, these embeddings are then passed through a 13-dimensional LFQ bottleneck (58), which effectively binarizes the representation. We read out the activations of this bottleneck as a 13-bit binary code which can be interpreted as one of $2^13 = 8,192$ discrete tokens (we note that all models will be updated to use 13-bit codes in the final paper version. The current CochStream versions besides the LibriSpeech version use 14-bit codes). Fourth, the output of the LFQ bottleneck is then projected to a 211 dimensional output, through two 1-dimensional convolutional layers (kernel size 10 and stride 1), separated by ReLU nonlinearities. This output corresponds to the frequencies in the cochleagram representation (7) which it is supervised to match via L2 error. Thus for every 5 seconds of audio, WavCoch extracts a sequence of 988 integers in the range [0, 8192) through the LFQ bottleneck, denoted as **cochlear tokens**, to feed into CochStream.

### 2.1.2 SEQUENCE MODELING: COCHSTREAM

**CochStream** is a GPT-style autoregressive Transformer (59). We train two versions: CochStream-base (97M parameters), with 12 layers, 12 attention heads and an embedding size of 784 and CochStream-large (1.3B parameters) with 24 layers, 16 attention heads, and an embedding size of 2,048. Both models have a vocabulary size of 16,384. The CochStream model takes as input the cochlear token sequence produced by WavCoch (Section 2.1.1) and predicts the next token in the sequence. The context length is approximately 20s (4,096 tokens). We utilize a learned positional embedding and compute the cross-entropy loss between the predicted logits and the true next token in the sequence.

Additionally we train an ablation model called CochStream-large-ll which is trained on the Libri-Light (60) dataset of 60k hours of publicly available speech recordings. This model has 1B parameters with an embedding size of 1024, 12 heads and 48 layers to match baseline models (22), (23). This model utilizes a WavCoch quantizer trained on the librispeech960 (61) dataset of 960 hours of publicly available speech, and utilizes a vocabulary size of 8,192, (following our ablations in Appendix Section 1.5).

### 2.1.3 TRAINING

To train WavCoch, we use the AdamW optimizer (62) with a peak learning rate of 1e-4 and a 200k step cosine-decay schedule. We use a batch size of 512, and a weight decay of 0.1. The training data consisted of internet audio clips containing naturalistic speech, largely podcasts and lecture recordings sampled at 16 kHz. We train the model on 500 hours of such data.

To train CochStream, we also use the AdamW optimizer with a peak learning rate of 3e-4 and a 200k step cosine-decay schedule with a 2k step warmup. We use weight decay of 0.1 and norm 1.0 gradient clipping. We train the model with a batch size of 256. CochStream was trained on 50,000 hours of such data (no associated transcriptions).

#### 2.1.4 OBTAINING COCHSTREAM EMBEDDINGS

We obtain CochStream embeddings by pooling the embeddings of all the tokens associated with the corresponding temporal section of the cochleagram via ground-truth phoneme or word boundaries (Section 3.1). For the pooling operation, we tested mean/max/min pooling for the linear probing experiments and lexical similarity for CochStream and the comparison models.

### 2.2 COMPARISON MODELS

CochStream is compared to three state-of-the-art speech representation models using the HuggingFace Transformers package (63): HuBERT-xl (identifier: *facebook/hubert-xlarge-ll60k*) (22), wav2vec2-large (identifier: *facebook/wav2vec2-large*) (64), wavLM (identifier: *microsoft/wavlm-base*, and *microsoft/wavlm-large*) (23). For the SUPERB benchmark, we additionally compare against two smaller models which share some similarity to our method, specifically, APC (26) and vq-wav2vec (65).

### 2.3 EVALUATION METRICS

To investigate the representational power of the CochStream embeddings, we linearly probe CochStream for phoneme identity, word identity, and lexical semantics. Additionally, to validate the use of CochStream as a powerful generic speech representation backbone, we evaluate it on the SUPERB benchmark (1), which fits specialized downstream task models on top of frozen backbone representations.

#### 2.3.1 PHONEME/WORD LINEAR PROBING

To probe for phoneme and word identity representation, we use the TIMIT dataset (66) consisting of approximately five hours of audio recordings with ground-truth phoneme- and word-boundaries. We use the train and complete test sets with exclusion of the "SA" sentences to allow for train and test sets that are non-overlapping in sentences as well as speakers. For phoneme classification, we followed the standard protocol of collapsing the TIMIT phoneme labels from 60 to 39 classes (67). The number of words in the TIMIT train set is 30,132 and 8,128 for the test set. The number of phonemes in the TIMIT train set is 140,225 and 50,754 in the test set. For linear probing, we use the scikit-learn LogisticRegression multiclass classifier (max_iter $= 10000$, solver $=$ lbfgs, penalty $=$ l2) (68). For each model, we identify the best-performing layer via mean pooling of the embeddings associated with each phoneme/word. The reported values are weighted accuracy scores (as the classes are imbalanced). To determine chance performance, we compute the probability of the most likely class label.

#### 2.3.2 LEXICAL SEMANTIC SIMILARITY (SSIMI)

We use the "sSIMI" lexical semantics benchmark developed for the ZeroSpeech 2021 challenge (69). The benchmark was constructed to probe whether audio-based models learn lexical semantics. The benchmark consists of pairs of words with ground-truth human similarity judgments (on a 0 and 10 scale) collected from behavioral experiments. For instance, a pair of words such as "water" and "river" have a human similarity score of 9.8, while a pair like "festival" and "whiskers" have a score of 0.2. Two audio subsets exist for these pairs of words: i) a natural dataset, consisting of the pairs of words present in LibriSpeech (70), and ii) a synthetic datasset, consisting of all pairs. The two datasets are each separated into a dev and test section (the LibriSpeech dataset is contains 309 and 3,753 word pairs for dev and test, and 705 and 9,744 for the synthesized subset). Each audio clip contains just one word, from which we extract embeddings. Following the procedure in (69), for each model, we identify the optimal embedding pooling strategy (mean/max/min) as well as layer based on the dev set, and obtain the final scores on the test set. The final sSIMI score is computed as

| Dataset | Model Size | Dataset Size | Phoneme Decoding | | Word Decoding | |
|---|---|---|---|---|---|---|
| | Params | Hours | Accuracy ↑ | Random ↑ | Accuracy ↑ | Random ↑ |
| HuBERT-xl | 1,000M | 60K | **0.93** | 0.20 | **0.88** | 0.07 |
| HuBERT-base | 97M | 1K | 0.85 | 0.20 | 0.77 | 0.07 |
| wav2vec2-large | 317M | 60K | 0.79 | 0.20 | 0.43 | 0.07 |
| WavLM-large | 317M | 94K | 0.91 | 0.20 | 0.85 | 0.07 |
| CochStream-base | 97M | 0.5K | 0.82 | 0.20 | 0.48 | 0.07 |
| CochStream-large | 1,300M | 50K | 0.92 | 0.20 | 0.67 | 0.07 |
| CochStream-large-ll | 1,000M | 60K | 0.92 | 0.20 | 0.69 | 0.07 |

Table 1: **Linear probing performance for phonemes or words on the TIMIT dataset**. Reported values are weighted accuracy scores on the TIMIT test set, consisting of non-overlapping sentences uttered by non-overlapping speakers from the train set.

the Spearman correlation between the cosine distance of embeddings for pairs of words and the true human similarity scores.

#### 2.3.3 Speech processing Universal PERformance Benchmark (SUPERB)

To comprehensively evaluate the strength of the representations produced by our model, we evaluate it on the SUPERB benchmark which contains 15 tasks, categorized into five aspects of speech: content, speaker, semantics, paralinguistics, and generation. We report values on a subset of 7 tasks spanning all 5 categories. We refer to the original paper for additional details on the benchmark (1).

## 3 Results

### 3.1 CochStream Embeddings Contain Information About Phoneme Identity, Word Identity, and Lexical Semantics

First, we investigate whether the representations from CochStream contain information about phoneme and word identity. The core premise is that if we can linearly decode these properties from the audio representations they will serve as useful representations for downstream tasks that require this information. We fitted linear classifiers on the phonemes/words in the train set of TIMIT (66) and tested the classifiers on the test set consisting of non-overlapping sentences and speakers (Section 2.3.1). We compared CochStream to three state-of-the-art models, HuBERT-xl (22), wav2vec2-large (64), and WavLM-large (23). As evidenced in Table 1, CochStream showed competitive performance compared to the comparison models. For phoneme decoding, CochStream surpassed wav2vec-large by a large margin and performed on par with HuBERT-xl and WavLM-large. The error patterns of CochStream were sensible. For instance, for phoneme probing, "er" was often confused with "r', or "ah" with "ih" (see Appendix I). For word decoding, CochStream once again surpassed wav2vec-large, however, CochStream fell short of HuBERT-xl and WavLM-large. We hypothesize that the subpar word decoding performance of CochStream relative to HuBERT and WavLM may be attributed to the the fact that HuBERT and its derivative models (WavLM) were exposed to global clustering operations aimed at discovering word-like units. In contrast, CochStream did not undergo any such global operations. Lastly, we emphasize the decoding performance for both phonemes and words scales well with CochStream size.

Second, we turn to a benchmark that goes beyond identity of phonemes/words, but instead asks whether the speech models learn representations of word meanings (lexical semantics). This benchmark (sSIMI; Section 2.3.2) evaluates the correlation between embeddings derived from audio associated with pairs of words (e.g., "water" and "river") to that of ground-truth human similarity judgments (Section 2.3.2) (69). The performance of speech models on this benchmark have been described as "modest" (69; 71), suggesting that this class of models struggle in learning semantics independently from the acoustic properties of words. Table 2 presents the performance of CochStream alongside the baseline models. The embeddings of CochStream showed superior performance on lexical semantics on both the natural and synthetic data subsets compared to the other models. Further,

| Dataset | Model Size | Dataset Size | LibriSpeech Audio | Synthetic Audio |
|---|---|---|---|---|
| | Params | Hours | Accuracy ↑ | Accuracy ↑ |
| HuBERT-xl | 1,000M | 60K | 7.81 | 10.37 |
| HuBERT-base | 97M | 1K | 6.10 | 7.48 |
| wav2vec2-large | 317M | 60K | 6.41 | 7.19 |
| WavLM-large | 317M | 60K | 10.50 | 10.41 |
| CochStream-base | 97M | 0.5K | 10.63 | 10.12 |
| CochStream-large | 1,300M | 50K | **12.52** | **10.64** |
| CochStream-large-ll | 1,000M | 60K | 10.99 | 10.52 |

Table 2: **Semantic similarity scores on the ZeroSpeech 2021 Lexical Semantic Benchmark**. Reported values are Spearman correlations (multiplied by 100 per (69)) between the embeddings for pairs of words versus human similarity judgments on the same pairs of words. The scores were obtained on the test sets of two dataset subsets (LibriSpeech and a synthetic set, see Section 2.3.2).

| Setting | Model | Dataset | PR | ASR | IC | KS | SID | ER | SS |
|---|---|---|---|---|---|---|---|---|---|
| | Params | Hours | PER ↓ | WER ↓ | Acc↑ | Acc↑ | Acc↑ | Acc↑ | SI-SDRi ↑ |
| HuBERT-large | 1,000M | 60K | 3.53 | 3.62 | 98.76 | 95.29 | 90.33 | 67.62 | 10.45 |
| wav2vec2-large | 317M | 60K | 4.75 | 3.75 | 95.28 | 96.66 | 86.14 | 65.64 | 10.02 |
| vq-wav2vec | 32M | 1K | 33.48 | 17.71 | 85.68 | 93.38 | 38.80 | 58.24 | 8.16 |
| WavLM-base | 97M | 1K | 4.84 | 6.21 | 98.42 | 96.79 | 84.51 | 65.94 | 10.37 |
| WavLM-large | 317M | 94K | 3.06 | 3.44 | 99.31 | 97.86 | 95.49 | 70.62 | 11.19 |
| APC | 4M | 0.36K | 41.98 | 21.28 | 74.69 | 91.01 | 60.42 | 59.33 | 8.92 |
| CochStream-base | 97M | 0.5K | 5.62 | 6.04 | 98.10 | 95.29 | 78.30 | 64.42 | 10.00 |
| CochStream-large | 1,300M | 50K | 4.20 | 3.87 | 98.10 | 96.12 | 82.14 | 68.37 | 10.67 |
| CochStream-large-ll | 1,000M | 60K | 4.16 | 3.78 | 97.85 | 95.72 | 80.02 | 68.98 | 10.66 |

Table 3: **Model embedding performance on various downstream tasks (SUPERB)**. Reported values are obtained by training a downstream task decoder on top of a frozen model backbone (1).

the CochStream model scores increased with model size. In sum, we demonstrate that CochStream embeddings learn state-of-the-art representations for lexical semantics.

### 3.2 Phoneme, Word, and Lexical Semantic Representations Across Time and Layers

We investigate how information about phonemes, words, and lexical semantics (Tables 1 and 2) is distributed across multiple layers in CochStream. In addition, we investigate the extent to which phoneme/word information is distributed across time. To do so, we run probing experiments on CochStream embeddings that are shifted in time (i.e., "intentionally" mismatched; Figure 2A bottom panel). Figure 2A shows that the highest accuracy on phoneme/word decoding was obtained at the true temporal window, with decreasing performance when the windows are shifted either left or right. Layer 1 had the lowest performance, with highest performance from middle layers. Figure 2B shows the lexical semantic benchmark, and in contrast to Figure 2A, the last layers of the model tend to exhibit the highest correlation with human similarity judgments.

### 3.3 CochStream Serves as a Strong Frozen Backbone for Downstream Audio Tasks

After having established that the CochStream representations themselves contain meaningful phoneme, word, and lexical semantics information (Section 3.1 and 3.2), we investigated whether the frozen representations of CochStream would serve as powerful features for training decoders across a range of various audio tasks. To do so, we leveraged 7 different SUPERB tasks, spanning the 5 broad categories present in this benchmark (1). As illustrated in 3 CochStream-large outperformed two of the most similar models evaluated on the benchmark – APC and vq-wav2vec, while performing competitively against state-of-the-art models on most other tasks. In particular, CochStream showed very strong performance on automatic speech recognition (SS), intent classification (IC), and speech

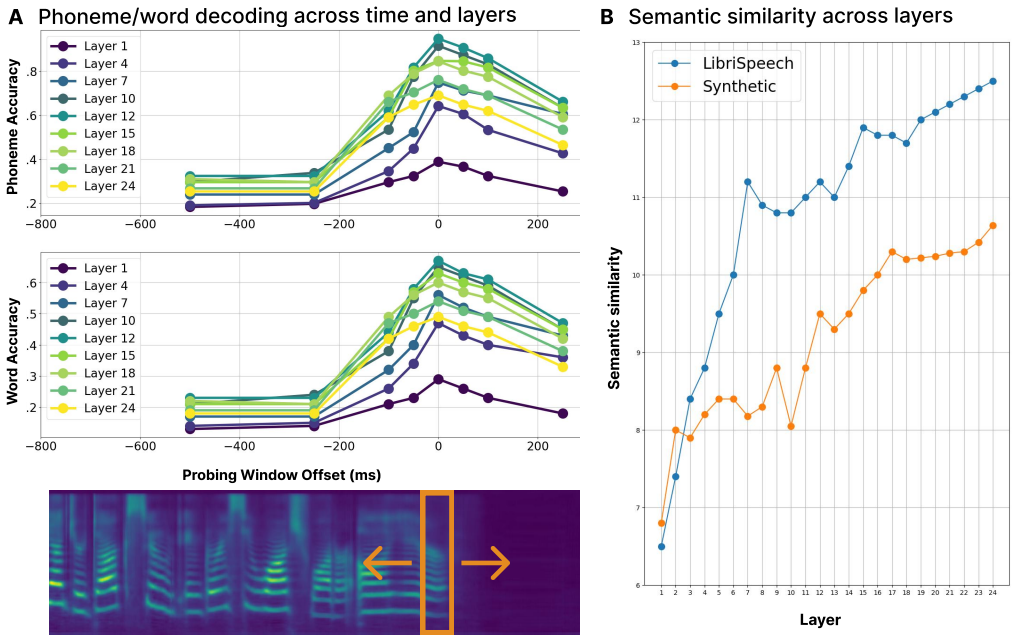


Figure 2: **A. Linear probing accuracy on TIMIT phonemes/words across time and layers.** The scores were obtained across model layers, with various shifts in temporal windows relative to the true phoneme/word from which the embeddings are extracted (x-axis) as demonstrated in the bottom panel. **B. Semantic similarity scores across layers on the ZeroSpeech 2021 benchmark.** Lines denote performance on the test set of the two data subsets.

separation (SS). In contrast, CochStream-large had subpar performance on speaker identification (SID) compared to the other models in the same parameter class. These findings illustrate that the features learned by CochStream serve as versatile representations for diverse downstream audio tasks, and importantly, demonstrate scalability (compare CochStream-base with CochStream-large).

### 3.4 CochStream Learns Short- and Long-Range Speech Statistics

Having established that CochStream representations contain decodable information of phonemes, words, and lexical semantics (Section 3.1), we ask: In the absence of ground-truth phoneme/word boundaries, does CochStream learn the statistics of speech? To answer this question, we leverage the fact that CochStream was trained to perform predictions in a space that can be visualized and interpreted, that is, the time-frequency cochleagram image (we note that models like HuBERT and wav2vec lack this capability). We hypothesize that if CochStream has learned the statistics of speech, it should manifest as two distinct modes: On short timescales, when provided enough conditioning (such as the first part of a common word), the model should be able to complete the cochleagram in a manner that is consistent with the remainder of the word. Conversely, at longer timescales, the model should diverge in the generated outputs, as many plausible words can follow any given phoneme or word. We provide CochStream with the beginnings of audio clips from the TIMIT test set (out-of-distribution for CochStream) and qualitatively analyze the resulting predictions (Figure 3).

To first test whether our model learns speech structure on short timescales, we provide CochStream with information about the first part of a common word (e.g., “she”, consisting of the phoneme “sh”), and evaluate its ability to predict the continuations from this first phoneme across different speakers in the TIMIT test set. As evidenced in Figure 3A, the model learns to consistently complete the phoneme with an “iy” phoneme, resulting in the word “she”. Conversely, if a phoneme has several likely continuations, the model learns to complete the phoneme with different words. As an example, we seed CochStream with the first phoneme (“wa”) of the word “water” and “wash” from two different speakers (Figure 3B). CochStream sometimes predicts the remainder of the true

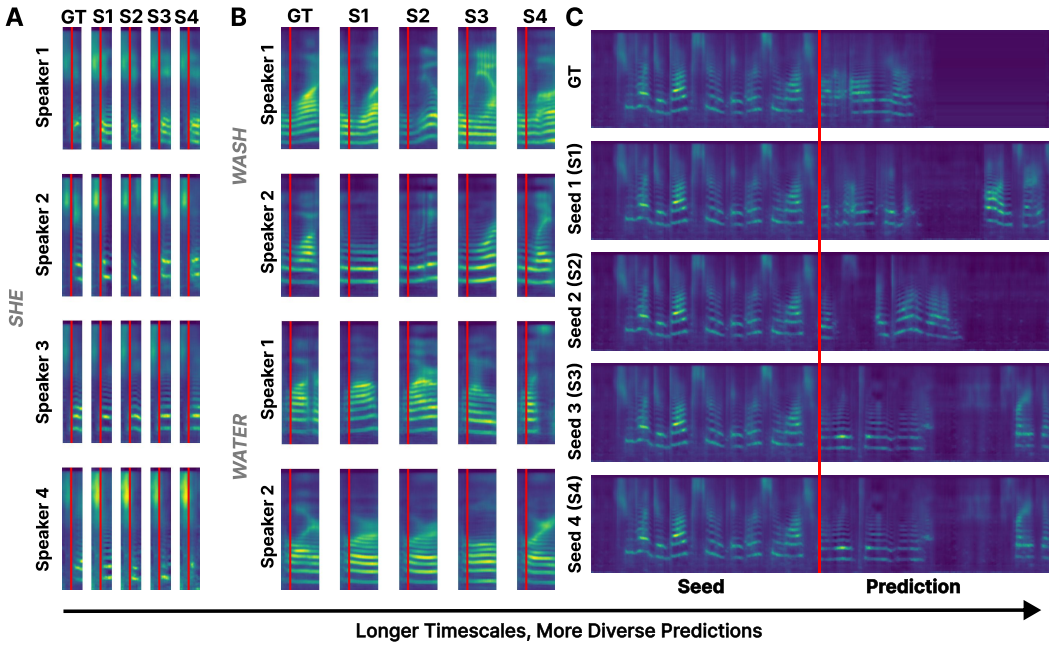


Figure 3: **Prediction of cochleagrams produced by the CochStream-base model: A.** CochStream is seeded with the first phoneme of the word "she" (left of red vertical line) and predicts the word completion (right of red line). The ground-truth (GT) cochleagram is shown in the first column.**B.** CochStream is seeded with the first phoneme of the words "wash" and "water". **C.** CochStream is seeded with the first 2.5 seconds of an audio clip (red vertical line) from the TIMIT test set and predicts the remaining part of the clip across four different seeds.

word, while sometimes it predicts a counterfactual word completion, which is still consistent with the first phoneme. For example, we observe that the prediction under Seed 3 of Speaker 1's utterance "wash" looks more similar to Speaker 2's ground-truth utterance of the word "water" than its own ground-truth word ("wash"). Hence, these visualizations suggest that CochStream learns the statistics of how phonemes compose to words.

Second, to test for diversity of longer-range predicted output, instead of simply seeding the model with a phoneme, we seed the model with the first 2.5 seconds of a TIMIT audio clip (Figure 3C). We observe that the model is capable of predicting several seconds of plausible completions. Furthermore, we highlight that various non-cherry-picked random seeds yield highly different predictions without collapsing. See additional examples here: https://anonymous.4open.science/w/cochstream-project-page-0546/.

## 4 Discussion and Limitations

In this paper, we proposed a self-supervised speech representation model, CochStream, for encoding speech without the need for any ground-truth text annotations.

We demonstrated that CochStream learns representations from which phonemes, word forms, and word meanings (lexical semantics) can be decoded at competitive levels, with state-of-the-art performance on lexical semantics (Section 3.1). Further, we demonstrated that CochStream serves as a strong representational backbone for various audio tasks, such as automatic speech recognition or speech separation (Section 3.3). For both types of tasks, we observed predictable increases in task performance as we scale the model. Finally, we demonstrate that our modeling framework is robust to the training dataset used (Section 3.3), emphasizing that the core performance contribution of our work comes from our novel two-stage modeling framework (WavCoch tokenization followed by auto-regressive CochStream prediction).

One strength of our framework is the use of cochlear tokens derived from a human-inspired waveform-to-cochleagram transformation (WavCoch), which serve as the input representations to a sequence model (CochStream). Taking a step back: biological systems have hierarchical stages that transform one known representation into another (known) representation. Here, we leverage this knowledge of representational states to train WavCoch and explicitly *probe this transformation* to extract cochlear tokens for downstream use. This approach–which we denote as *"Transformation Imitation"*–holds great potential for extension to other perceptual domains. Another strength of cochlear tokens is their efficiency: each second of audio is represented with just 197 tokens. This downsampling makes the input inherently efficient, enabling real-time predictions. Importantly, our Transformation Imitation framework is adaptable to different representations; for example, the cochleagram could easily be replaced with a standard mel-spectrogram. While we do not make explicit claims about the superiority of the cochleagram representation over the mel-spectrogram, our exploratory findings indicate that cochleagrams perform at least as well as spectrograms (as estimated through codebook usage and phoneme purity, see Appendix Section 1.5). Finally, we acknowledge that previous work has leveraged cochleagrams in speech representation learning (27; 72; 7) and various neural codec approaches are similar in spirit to WavCoch in the fact that they extract intermediate bottleneck representations for downstream tasks (8; 11; 12; 13; 14; 15; 16). However, we still believe that WavCoch is novel in its way of learning to transform one representation into *another* representation through a discrete quantization bottleneck (instead of auto-encoding, as done in related approaches).

A final advantage to emphasize is that CochStream allows to visualize and interpret resulting audio predictions directly (for example, Figure 3), a capability that many audio-based models do not have, making CochStream less of a "black box". Although the predicted outputs of CochStream cannot directly be transformed back into audio, we have performed initial analyses to generate audible responses see Appendix Section 1.6).

A couple limitations of our work exist. One limitation is that our model is trained on English speech, constraining the analyses we perform to tasks and materials in English (73; 74). We do emphasize that our model, as well as textless natural language processing more broadly, can serve as particularly useful for applications within low-resource languages. Another limitation is that due compute and time constraints, our models could not be scaled to their full potential.

In addition to CochStream's advantages for machine learning, our hope is that it will serve as a valuable model for the emerging field of "NeuroAI" (75). This field leverages artificial neural networks to better understand human behavior and brain function, testing whether artificial models trained on naturalistic data and/or biologically relevant tasks develop human-like solutions. Recent work has explored whether these models mimic human behavioral characteristics (44; 40; 45; 46; 47), as well as whether their internal representations align with neural activity in the human brain (48; 49; 50; 51; 52; 76).

We by no means claim that CochStream is a perfect biologically-inspired model, but it is a critical step in the right direction. CochStream uses a Transformer architecture—a choice driven by its proven efficacy for sequence-based tasks—which might have properties that make it incompatible with biological neurons (however, see 77; 78; 79 for how Transformer mechanisms could be implemented in biological hardware). Nevertheless, CochStream marks a significant step forward: it leverages biologically-realistic input representations and employs a simple, causal prediction mechanism without relying on additional hand-crafted heuristics. This approach contrasts with many speech representation models that depend on high-fidelity signal reconstruction objectives, bidirectional non-causal prediction, and the contrasting of hundreds of samples within a training batch (see Introduction 1).

In conclusion, we present a novel framework for speech representation learning, aiming to advance the development of more human-like, neurally plausible models that capture the intricacies of human speech processing.

## REFERENCES


[1] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[2] Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kaiwei Chang, Jiawei Du, Ke-Han Lu, Alexander H Liu, Ho-Lam Chung, Yuan-Kuei Wu, Dongchao Yang, et al. Codec-superb@ slt 2024: A lightweight benchmark for neural audio codec models. *arXiv preprint arXiv:2409.14085*, 2024.

[3] Ambuj Mehrish, Navonil Majumder, Rishabh Bhardwaj, Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing, 2023.

[4] Kuansan Wang and Shihab Shamma. Self-normalization and noise-robustness in early auditory representations. *IEEE transactions on speech and audio processing*, 2(3):421–435, 1994.

[5] Kevin N. Ochsner and Stephen Kosslyn. *The Oxford Handbook of Cognitive Neuroscience, Volume 1: Core Topics*. Oxford University Press, 12 2013.

[6] Taishih Chi, Powen Ru, and Shihab A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, August 2005.

[7] Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023.

[8] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[9] Vladimir E. Iashin and Esa Rahtu. Taming visually guided sound generation. *ArXiv*, abs/2110.08791, 2021.

[10] Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022.

[11] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

[12] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Shengpeng Ji, Minghui Fang, Ziyue Jiang, Rongjie Huang, Jialung Zuo, Shulei Wang, and Zhou Zhao. Language-codec: Reducing the gaps between discrete codec representation and speech language models. *arXiv preprint arXiv:2402.12208*, 2024.

[14] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. *arXiv preprint arXiv:2404.02781*, 2024.

[16] Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Boris Ginsburg. Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis. *arXiv preprint arXiv:2406.05298*, 2024.

[17] Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 591–595. IEEE, 2024.

[18] Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *ArXiv*, abs/2301.02111, 2023.

[19] Matt Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. *ArXiv*, abs/2306.15687, 2023.

[20] Kayoko Okada, Feng Rong, Jon Venezia, William Matchin, I.-Hui Hsieh, Kourosh Saberi, John T. Serences, and Gregory Hickok. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex (New York, N.Y.: 1991)*, 20(10):2486–2495, October 2010.

[21] Claire Tang, LS Hamilton, and EF Chang. Intonational speech prosody encoding in the human auditory cortex. *Science*, 357(6353):797–801, 2017.

[22] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[23] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[24] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training, September 2021. arXiv:2108.06209 [cs, eess].

[25] Paul Michel, Okko Räsänen, Roland Thiolliere, and Emmanuel Dupoux. Blind phoneme segmentation with temporal prediction errors. *arXiv preprint arXiv:1608.00508*, 2016.

[26] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*, 2019.

[27] Cory Shain and Micha Elsner. Acquiring language from speech by learning to remember and predict. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 195–214, 2020.

[28] Yu-An Chung, Hao Tang, and James Glass. Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*, 2020.

[29] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE, 2021.

[30] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020.

[31] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

[32] Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*, 2021.

[33] Guan-Wei Wu, Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. Improving textless spoken language understanding with discrete units as intermediate target. *arXiv preprint arXiv:2305.18096*, 2023.

[34] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, June 2019.

[36] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, December 2020.

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[38] Saurabhchand Bhati, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. Unsupervised speech segmentation and variable rate representation learning using segmental contrastive predictive coding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2002–2014, 2022.

[39] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879, 2020.

[40] Mark R. Saddler, Ray Gonzalez, and Josh H. McDermott. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, 12(1):7278, December 2021.

[41] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020. arXiv: 2005.14165.

[42] Sidney R. Lehky and Terrence J. Sejnowski. Network model of shape-from-shading: neural function arises from both receptive and projective fields. *Nature*, 333(6172):452–454, June 1988. Number: 6172 Publisher: Nature Publishing Group.

[43] Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10, September 2016.

[44] Thomas Schatz, Naomi H Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7):e2001844118, 2021.

[45] Juliette Millet and Ewan Dunbar. Do self-supervised speech models develop human-like perception biases? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7591–7605, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[46] Andrew Francl and Josh H. McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, 6(1):111–133, January 2022.

[47] Marianne de Heer Kloots and Willem Zuidema. Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2. 0. In *Proc. INTERSPEECH*, 2024.

[48] Alexander J. E. Kell, Daniel L.K. 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.e16, May 2018.

[49] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, June 2022.

[50] Yuanning Li, Gopala K. Anumanchipalli, Abdelrahman Mohamed, Junfeng Lu, Jinsong Wu, and Edward F. Chang. Dissecting neural computations of the human auditory pathway using deep neural networks for speech. *bioRxiv*, 2022. Publisher: Cold Spring Harbor Laboratory.

[51] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *bioRxiv*, 2023. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2023/06/14/2022.09.06.506680.full.pdf.

[52] Subba Reddy Oota, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *INTERSPEECH 2023-24th INTERSPEECH Conference*, pages 5167–5171, 2023.

[53] Friedemann Pulvermüller, Rosario Tomasello, Malte R Henningsen-Schomers, and Thomas Wennekers. Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22(8):488–502, 2021.

[54] Federico Adolfi, Jeffrey S. Bowers, and David Poeppel. Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Networks*, 162:199–211, May 2023. arXiv:2204.03740 [cs, eess, q-bio].

[55] Brian R Glasberg and Brian C. J Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103–138, August 1990.

[56] Josh H. McDermott and Eero P. Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, September 2011.

[57] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[58] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David C. Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. 2023.

[59] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.

[60] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazar'e, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdel rahman Mohamed, and Emmanuel Dupoux. Libri-light: A benchmark for asr with limited or no supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2019.

[61] 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, April 2015. ISSN: 2379-190X.

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[63] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: state-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[64] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[65] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *ArXiv*, abs/1910.05453, 2019.

[66] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[67] K-F Lee and H-W Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

[68] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[69] Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588*, 2020.

[70] 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.

[71] Jaeyeon Kim, Injune Hwang, and Kyogu Lee. Learning semantic information from raw audio signal using both contextual and phonetic representations, 2024.

[72] Maren Buermann and TA Van Meer. Speech recognition using very deep neural networks: Spectrograms vs cochleagrams. *Tom van Meer's Lab, Eindhoven Univ. Technol., Eindhoven, The Netherlands, Tech. Rep*, 2020.

[73] Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170, 2022.

[74] Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567, 2024.

[75] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, James DiCarlo, Surya Ganguli, Jeff Hawkins, Konrad Körding, Alexei Koulakov, Yann LeCun, Timothy Lillicrap, Adam Marblestone, Bruno Olshausen, Alexandre Pouget, Cristina Savin, Terrence Sejnowski, Eero Simoncelli, Sara Solla, David Sussillo, Andreas S. Tolias, and Doris Tsao. Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nature Communications*, 14(1):1597, March 2023. Number: 1 Publisher: Nature Publishing Group.

[76] Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. *arXiv preprint arXiv:2410.09230*, 2024.

[77] Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. *Advances in Neural Information Processing Systems*, 34:15301–15315, 2021.

[78] James C. R. Whittington, Joseph Warren, and Timothy E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation, 2022.

[79] Leo Kozachkov, Ksenia V. Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34):e2219150120, August 2023. Publisher: Proceedings of the National Academy of Sciences.