LibriBrain100: One Hundred Hours of Broad and Deep MEG Data for Neural Speech Decoding at Scale

Francesco Mantegna[1] **Dulhan Jayalath**[1] **Gereon Elvers**[1] **Tasha Kim**[1]

Benjamin Ballyk[1] **Alex Fung**[1,2] **SungJun Cho**[1,3] **Teyun Kwon**[1]

Luisa Kurth[1] **Miran Özdogan**[1] **Gilad Landau**[1] **Pratik Somaiya**[1]

Natalie Voets[2] **Mark Woolrich**[3] **Oiwi Parker Jones**[1]

[1]PNPL🍍, Department of Engineering Science, University of Oxford, UK
[2]FMRIB, Oxford Centre for Integrative Neuroimaging, University of Oxford, UK
[3]OHBA, Oxford Centre for Integrative Neuroimaging, University of Oxford, UK

{francesco, oiwi}@robots.ox.ac.uk

Abstract

We introduce LibriBrain100, a large-scale MEG dataset for speech decoding designed from the ground up for reproducible, standardised evaluation. LibriBrain100 more than doubles the size of the original LibriBrain release, resulting in **over 100 hours of high-quality MEG** acquired while subjects listened to naturalistic continuous speech. With ∼**80 hours from a single subject**, LibriBrain100 sets a new record for deep, within-subject neural data (8× more than the next comparable dataset and roughly 80× more than other datasets). To demonstrate the payoff of this depth-first design, we evaluate on a word-classification benchmark—an increasingly well-established stepping stone towards the open challenge of non-invasive brain-to-text decoding. Using an existing decoding model, we achieve state-of-the-art performance—validating both the quality of the recordings and the value of within-subject data at scale. Because collecting 80 hours of data per user is impractical for real-world applications, we also collected ∼**40 minutes of additional data from each of 32 subjects**. Using the same word-classification benchmark, we demonstrate the value of broad multi-subject data: supervised fine-tuning of a pre-trained model can substantially compensate for limited per-subject data. We provide **standard train, validation, and test splits, all reproducible through an open-sourced Python library** that supports easy downloading, optional preprocessing, and data loading for common deep learning frameworks. In addition, the dataset and evaluation infrastructure are being released alongside an open machine-learning competition with a public leaderboard for standardised benchmarking. Ultimately, our hope is that LibriBrain100 will accelerate progress towards practical non-invasive brain-computer interfaces, capable of restoring communication to people living with severe paralysis.

Preprint.

1 Introduction

Deep learning had a transformational effect on computer vision once ImageNet gave it a benchmark: data at scale and standard evaluation protocols. Neural speech decoding is at a similar inflection point — but with an added complexity. In invasive settings, systems decoding from surgically implanted electrodes now outperform automatic speech recognition in some conditions, achieving word error rates below 5% in paralysed patients (Card et al., 2024; Willett et al., 2023). But invasive recording requires brain surgery, limiting deployment to clinical trials and excluding the vast majority of people who might benefit. Non-invasive alternatives are therefore the real prize, and deep learning is beginning to show real gains there too (Tang et al., 2023; Défossez et al., 2023; d'Ascoli et al., 2025; Jayalath et al., 2025a). Yet unlike computer vision, the field has lacked the shared infrastructure to know how much progress is real, what can be built on reliably, and how quickly to anticipate advances. Without standard benchmarks, shared data, and common evaluation protocols, it is difficult to know what works or by how much.

The LibriBrain dataset (Özdogan et al., 2025) — and its open ML competition (Landau et al., 2025) — represented a significant step in the right direction. Taking the insight that within-subject data pushes results fastest (d'Ascoli et al., 2025), LibriBrain provided the largest within-subject magnetoencephalography (MEG) dataset for speech decoding from brain recordings to date, together with standard splits and a curriculum of benchmark tasks. Building on the LibriBrain infrastructure, subsequent work has advanced the state of the art across speech detection, phoneme classification, word classification, and even brain-to-text (Elvers et al., 2026; Jayalath et al., 2025a; Jayalath and Parker Jones, 2026).

But LibriBrain has its limitations. It covers only a single subject listening to a single genre (detective fiction) and so does not speak directly to cross-subject generalisation. This is of a critical importance for real-world BCIs, where per-user data collection must be minimal. The limited stimulus diversity inherent in a tight focus on the series of Sherlock Holmes books, leaves nuanced questions about phonetic and semantic decoding underexplored. Finally, even with its unprecedented scale of within-subject data, analyses of LibriBrain have shown no signs of plateauing (see, e.g., Özdogan et al., 2025), suggesting that more data should yield even bigger gains.

LibriBrain100 addresses these limitations directly. It extends LibriBrain both in *depth* and in *breadth*. Along the depth axis, LibriBrain100 extends the within-subject data from $\sim$50 to $\sim$80 hours, completing the entire canon of Sherlock Holmes and adding new stimuli designed with both phonetics and semantics in mind: TIMIT (Garofolo et al., 1993b) and MOCHA-TIMIT (Wrench, 1999, 2000) are standard corpora in both ASR research (e.g., Mohamed et al., 2009; Hinton et al., 2012; Graves et al., 2013) and invasive speech BCIs (e.g., Mesgarani et al., 2014; Cheung et al., 2016; Anumanchipalli et al., 2019; Makin et al., 2020; Metzger et al., 2023), which control for phoneme and phoneme-pair distributions and enable direct comparison with prior studies using the same stimuli; and podcast narratives spanning a wide range of semantic domains, which were previously used for fMRI-based semantic decoding (Tang et al., 2023) enabling cross-modal comparison. Along the breadth axis, LibriBrain100 extends data collection to 32 new subjects, each contributing $\sim$40 minutes of MEG, designed to support supervised fine-tuning and data-efficient cross-subject generalisation. Together, these corpora span the acoustic–semantic axis, positioning LibriBrain100 to address the questions of how MEG-based speech decoding may be driven by representations of sound versus meaning in the brain.

On the evaluation side, LibriBrain100 advances the curriculum of tasks from speech detection and phoneme classification to word classification — a task capable of utilising both phonetic and semantic representations in the brain, and that sits closer to the brain-to-text goal that motivates much of the field (Figure 1). We evaluate word detection in two settings: within-subject, targeting state-of-the-art performance by scaling up training data for subject 0; and cross-subject, focusing on data-efficient generalisation via supervised fine-tuning from a single high-data subject to new subjects with limited recordings. Standard splits are provided across all sub-datasets, and baseline results using MEG-XL (Jayalath and Parker Jones, 2026) are released to support reproducible community benchmarking.

LibriBrain100 makes the following contributions:

- **A large-scale, stimulus-diverse within-subject MEG dataset**: $\sim$80 hours of recordings from subject 0 — the largest within-subject speech decoding dataset to date — completing the full Sherlock Holmes canon and adding TIMIT and MOCHA-TIMIT for phonetically

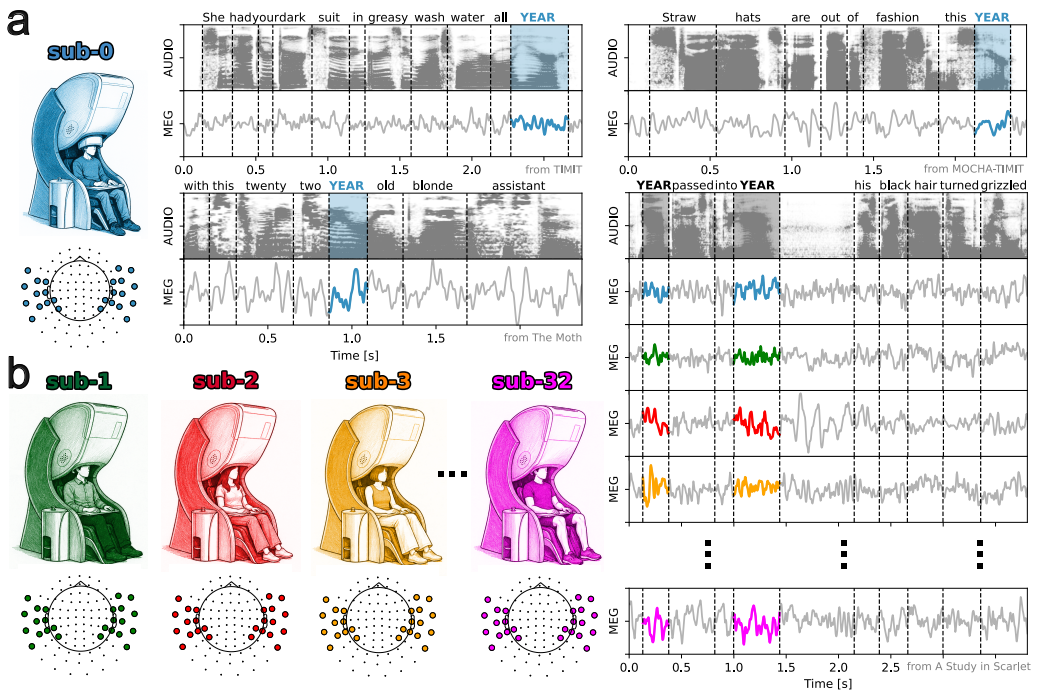

Figure 1: **Decoding words from LibriBrain100**. LibriBrain100 supports word-classification benchmarks across both deep within-subject data and broad cross-subject data. (**a**) *Within-subject decoding*: example speech stimuli and MEG responses from subject 0 across multiple corpora. Target words are highlighted in the text, aligned to the acoustic spectrogram, and marked in the corresponding MEG time series. MEG time series were averaged across temporal sensors illustrated in the sensor layout. (**b**) *Cross-subject decoding*: the same target-word classification task can be evaluated across the 32 additional subjects with limited per-subject data, enabling benchmarking for data-efficient generalisation.

controlled analyses and podcast narratives for semantic decoding — enabling cross-modal comparison with the fMRI literature and direct comparison with influential invasive BCI studies

- **A multi-subject MEG dataset for cross-subject generalisation**: recordings from 32 new subjects ($\sim$40 minutes each), designed to support supervised fine-tuning and data-efficient generalisation benchmarking
- **A word classification benchmark with standard splits and baselines**: an extended evaluation curriculum with train/validation/test splits across all sub-datasets, together with reproducible baseline results for within-subject and cross-subject settings using MEG-XL
- **Open data release and tooling**: raw and preprocessed data on HuggingFace, rich linguistic annotations spanning sub-lexical to lexical levels, and a Python library for streamlined deep learning integration — all resources to support a planned open ML competition

2 Related Work

In recent decoding work on MEG acquired during naturalistic speech listening (e.g., Défossez et al., 2023; d'Ascoli et al., 2025; Jayalath et al., 2025a,b; Jayalath and Parker Jones, 2026), a recurring set of large-scale MEG datasets have emerged (also see King et al., 2026). The usual suspects include MAUS (Schoffelen et al., 2019), MEG-MASC (Gwilliams et al., 2023), Armeni et al. (2022), Le Petit Prince (Bel et al., 2026), and LibriBrain (Özdogan et al., 2025). Table 2 (Appendix A.2) summarises these datasets along with LibriBrain100, listing details such as the language of stimuli and dimensions in which we can measure their size. In terms of 'deep', within-subject data, LibriBrain100 is $\sim 8\times$ bigger than the next non-LibriBrain dataset (Armeni et al., 2022), and $\sim 80\times$ bigger than the rest.

We can compare the existing datasets in terms of *breadth*—the number of subjects—and *depth*—the hours of data per subject. On the one hand, MAUS, MEG-MASC, and Le Petit Prince represent breadth-first designs, spanning 27–96 subjects but with under 2 hours per subject. Armeni and LibriBrain contrast depth-first designs, with 10 and 52.3 hours per subject respectively, but only 1–3 subjects. This distinction matters for decoding. Despite other datasets having greater total hours, d'Ascoli et al. (2025) find word-classification performance to be strongest by a substantial margin for Armeni—the deepest dataset in their study, LibriBrain having not come out yet. When LibriBrain is included, as in Jayalath et al. 2025a, decoding performance is even stronger with it than with Armeni, further underscoring the importance of within-subject depth.

The obvious trade-off for LibriBrain, which is the closest dataset to LibriBrain100, is that its depth comes at the cost of any breadth—the original LibriBrain release contained just one subject. LibriBrain100 addresses this limitation while continuing to prioritise depth. It deepens the single-subject recording to ∼80 hours while adding ∼40 minutes of data from each of 32 additional subjects—resulting in a hrs/subject range of 0.6–80.0. This reflects both a depth-first design and respectable provisions for studying generalisation to new subjects with more realistic amounts of data—of the six datasets in Table 2, LibriBrain100 has the 3rd highest number of subjects. The nature of the audio stimuli in these datasets is also interesting. Unlike LibriBrain, which focused only on readings of Sherlock Holmes (seven out of nine books), LibriBrain100's stimulus set—which includes TIMIT (Garofolo et al., 1993b), MOCHA-TIMIT (Wrench, 1999, 2000), and 30 podcasts-from *The Moth* (https://themoth.org)–is designed to better control for sound and meaning-based brain activity.

3 The LibriBrain100 Dataset

3.1 Overview

The LibriBrain100 dataset contains over 100 hours of non-invasive magnetoencephalography (MEG) recordings, which were acquired while subjects listened to connected speech (Table 1). In addition, the dataset includes paired event files with time-locked annotations for linguistic events (e.g. speech/non-speech, phonemes, words). These annotations were designed for use as labels for decoding tasks such as word classification (Section 4). The MEG data are being released in both raw and preprocessed/serialised formats, where both may be thought of as matrices with 306 sensor dimensions $\times$ T time samples. MEG recordings were originally sampled at 1 kHz but then downsampled to 250 Hz during preprocessing to preserve oscillations into the high-gamma range (70–125 Hz). Consequently, each sample $t \in [1, T]$ represents 1 ms in the raw data and 4 ms in the preprocessed data.

Table 1: **LibriBrain100 data sources and standard splits**. The dataset comprises a *deep* single-subject component (∼80 hours) and a *broad* multi-subject component (∼40 minutes per subject across 32 subjects). All durations are shown in hours and rounded to one decimal place; totals are computed before rounding. For a visual summary of the dataset, see Figure 5 (Appendix A.1).

Component	Source	# Subj	All Data	Train	Val	Test
Deep						
Subject 0	Sherlock (all books)	1	68.1	67.3	0.4	0.4
	TIMIT	1	5.3	4.9	0.2	0.2
	MOCHA-TIMIT	1	1.1	0.8	0.2	0.2
	The Moth (30 podcasts)	1	6.0	5.7	0.2	0.2
Subtotal (hours)		1	80.5	78.6	1.0	0.9
Broad						
Subjects 1–32	Sherlock (2 chapters) (∼44 minutes/subject)	32	23.7	–	11.5	12.2
Total (hours)		**33**	**104.2**	**78.6**	**12.5**	**13.1**

3.2 Data Collection and Structure

All MEG data were collected at the Oxford Centre for Human Brain Activity (OHBA) on a MEGIN TRIUX™ Neo system with 306 sensors known as superconducting quantum interference devices

(SQUIDs) (Hämäläinen et al., 1993). The SQUIDs in this system partition into 102 magnetometers and 204 gradiometers. In a magnetically shielded room, these sensors are sufficiently sensitive to detect electromagnetic fields generated by cortical activity throughout the whole brain (Bénar et al., 2021). The present release includes data from 33 subjects, all of whom were proficient English speakers and provided informed consent for pseudonymised data to be shared; this study was approved by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (R90053/RE003).

In terms of structure, the majority of the data (∼80 hours) comes from a single subject, referred to as 'subject zero' (`sub-0`). This was a deliberate choice, as hours of data per subject is known to drive performance more than total number of hours across subjects (d'Ascoli et al., 2025). At the same time, for downstream applications where it would be preferable not to require such large data from every new subject, we also include more practical amounts of data (∼20 minutes × 2 sessions per subject) from a cohort of 32 additional subjects (`sub-1` to `sub-32`). In total, the dataset includes more than 100 hours of MEG data with annotations generated from the audio stimuli that the subjects listened to.

The audio stimuli include LibriVox recordings for every book in the canon of Sherlock Holmes (Doyle, 1887, 1890, 1892, 1893, 1902, 1905, 1915, 1917, 1927). While the first seven of these books (∼50 hours) featured in the original LibriBrain dataset (Özdogan et al., 2025), completing the canon here was not only completionist but aimed to maximise data quantity while minimising variability. For example, all but one book were read by the same narrator and both narrators selected to have similar Southern British accents (see Appendix B.2 for more details). Our aim in limiting the variability of these stimuli was pragmatic. By analogy, if you wanted to train automatic speech recognition (ASR) from scratch on about 68 hours of audio recordings, it would be easier to concentrate on just one or two speakers than attempt to model every accent or dialect.

To complement the Sherlock Holmes audiobooks, we looked to the literature for inspiration. For better phonetic coverage, we used TIMIT (Garofolo et al., 1993b) and MOCHA-TIMIT (Wrench, 1999, 2000). Not only has TIMIT long been a cornerstone for the development of ASR technology (e.g., Graves et al., 2013), but it has played a foundational role in our understanding of speech comprehension in the brain (e.g., Mesgarani et al., 2014; Cheung et al., 2016). One feature of TIMIT is that the sentences in it were designed to balance the distribution of phonemes and phoneme-bigrams. Unlike Sherlock, TIMIT also contains audio recordings from 630 speakers across 8 dialect groups in the USA; MOCHA-TIMIT adds male and female speakers from the UK (see Appendix B.2).

For semantic coverage, we looked to a landmark semantic decoding study (Tang et al., 2023) which used multiple short (∼12 minute) podcast stories to cover a broad range of topics. We include 30 podcast stories from *The Moth* (for details, see Appendix B.2).

For reproducible evaluation, the data are split into standard training, validation, and testing sets. Following both Özdogan et al. (2025) and Landau et al. (2025), sessions 11 and 12—which correspond to chapters 11 and 12—from the first Sherlock book are held out for validation and testing, respectively. Sessions 13 and 14, which were previously held-out for competition evaluation (Landau et al., 2025), have now made public and can be used for training, together with the rest of the Sherlock data. As there is a standard split for TIMIT, we use that in the MEG data for comparability. Concretely, we use the 24-speaker core test set, which consists of 2 male and 1 female speaker from each dialect region, totalling 192 utterances (24 speakers × 8 utterances which exclude the shared `SA1` and `SA2` sentences). For validation, we follow the standard Kaldi TIMIT recipe and use its 50-speaker development set, again excluding the shared `SA1` and `SA2` sentences. Although surgical decoding studies have used MOCHA-TIMIT (e.g., Anumanchipalli et al., 2019; Makin et al., 2020), we are unaware of a precedent for a standard split; therefore, we created our own split. To do this, we note that MOCHA-TIMIT is organised into four sets, A–D, where sentences repeat between sets A and D and between sets B and C. We use A and D for training. For validation and testing, we split the sentences in sets B and C into equally sized subsets of unique sentences. This results in unique sentences for train, validation, and test sets, minimising the risk of information leakage due to sentence type. Table 1 summarises the data and splits.

3.3 Data Formats, Access, and Supporting Python Library

We are releasing LibriBrain100 in two formats. First, we are releasing the raw data (with no preprocessing) in BIDS format (Niso et al., 2018). BIDS structures are directories that follow guidelines on file naming and organisation, with MEG data in FIF format. Technically, LibriBrain100 is a collection of BIDS datasets, one for each Sherlock book, TIMIT, MOCHA-TIMIT, and The Moth podcasts. Second, we are releasing the data in a ready serialised format. These data have been minimally preprocessed (i.e., corrected for head movement, filtered, downsampled) and then converted to float32 HDF5 format, for easy machine learning.

The MEG data in both FIF and HDF5 files can be thought of as a 306 channel $\times$ T time-sample matrix, with time sampled at 1,000 Hz for the raw data or 250 Hz for the serialised data. Each FIF or HDF5 file is paired with an annotation file in TSV format. The annotation files contain time-stamps for linguistic events like word onsets, which can be used to define labels for supervised learning.

The recommended way to access the data is through our custom Python library, `pnpl` (`pip install pnpl`). The library provides task-driven `torch.utils.data.Dataset` classes whose constructors return ready-to-iterate windows of MEG, time-locked to events such as word or phoneme onsets:

```python
from pnpl.datasets import LibriBrain100
from pnpl.tasks import WordClassification

ds = LibriBrain100(
    data_path="./data/LibriBrain100",
    task=WordClassification(tmin=0.2, tmax=0.6),
    partition="train",
)

x, y = ds[0]   # x: (channels, time), y: integer word class id
```

If the requested files are not at `data_path`, the library downloads them on demand from Hugging Face (the `LibriBrain100` loader transparently fetches each record from whichever underlying repository owns it). Given the size of the full dataset ($\sim$104 hours of MEG, hundreds of GB even after serialisation), the constructor exposes selectors so users can scope the download to the part they care about:

```python
# Subject 0 only -- the deep component (~80 hours):
ds = LibriBrain100(data_path=..., task=..., partition="train",
                   subjects="deep")

# Only TIMIT, only on subject 0:
ds = LibriBrain100(data_path=..., task=...,
                   subjects="deep", corpus="timit")

# Multiple corpora at once, still only on subject 0:
ds = LibriBrain100(data_path=..., task=...,
                   subjects="deep", corpus=["timit", "mocha"])
```

The `subjects=` argument accepts `"all"` (default), `"deep"` (subject 0), `"broad"` (subjects 1–32), an integer, a single subject id (`"0"` or `"sub-0"`), or any iterable of those. The `corpus=` argument accepts `"all"` (default), `"sherlock"`, `"timit"`, `"mocha"`, `"podcasts"`, or a list of those (the canonical token for "MOCHA-TIMIT" is `"mocha"` but aliases like `"mocha-timit"` or `"the_moth"` are accepted). As the broad component (subjects 1–32) was only collected with the Sherlock stimuli, the loader rejects combinations like `subjects="broad", corpus="timit"` up front rather than silently returning an empty or incomplete dataset. Per-task wrappers (`LibriBrain100Speech`, `LibriBrain100Phoneme`, `LibriBrain100Word`) are also available for convenience. For finer control, users can pass a list of explicit `include_run_keys` or `exclude_run_keys` as `(subject, session, task, run)` tuples, mirroring the run-key convention from the original LibriBrain release.

While we recommend accessing the data through the Python library, with its high-level API which does not require an understanding of the underlying repository structure, the data are also available for direct download from two Hugging Face repositories: the original LibriBrain repository (https:

//huggingface.co/datasets/pnpl/LibriBrain) and a new LibriBrain2 repository (https://huggingface.co/datasets/pnpl/LibriBrain2). For philosophers and fans of category mistakes (Ryle, 1949), we might point out that LibriBrain100 is not a third repository alongside LibriBrain and LibriBrain2, but rather their union — as the University of Oxford is not a building alongside its colleges, departments, and other facilities, but rather the collection of them all.

4 Decoding Experiments

To evaluate LibriBrain100, we focus here on the task of word classification which is defined as follows. Given a multi-channel MEG recording $\mathbf{X} \in \mathbb{R}^{C \times T}$ with C channels and T time samples, the word classification task maps to a target word $y \in \mathcal{V}$ where $\mathcal{V}$ is a fixed vocabulary. Word classification has previously been studied in MEG using the most frequent 50 and 250 words (d'Ascoli et al., 2025; Jayalath et al., 2025a; Jayalath and Parker Jones, 2026) defined over various corpora and using top-10 balanced accuracy, computed by averaging per-class top-10 accuracy across the vocabulary $\mathcal{V}$ to account for imbalances in word frequency (Zipf, 1949). In the following experiments, we use a vocabulary of 50 words (see Appendix F.2).

As a backbone, we use MEG-XL (Jayalath and Parker Jones, 2026). This pre-trained model has achieved state-of-the-art word classification results in cross-subject MEG decoding using unsupervised pre-training followed by supervised fine-tuning. MEG-XL draws on insights from LLMs to significantly improve downstream performance by extending the temporal context (5–300×) of previous foundation models of the brain (cf. Jiang et al., 2024; Avramidis et al., 2025; Wang et al., 2024; Xiao et al., 2025; Wang et al., 2025). MEG-XL has been pre-trained on approximately 300 hours from 800 subjects over multiple speech and non-speech datasets—i.e., MOUS (Schoffelen et al., 2019), Cam-CAN (Shafto et al., 2014), and SMN4Lang (Wang et al., 2022)—before being fine-tuned on LibriBrain (Özdogan et al., 2025). In this paper, we fine-tune the base MEG-XL checkpoint (https://huggingface.co/pnpl/MEG-XL) on LibriBrain100 for the following tasks.

4.1 Task 1: From Many-to-One — Improving Within-Subject Word Classification

Figure 2 describes the results on the held-out test data of Subject 0 when fine-tuning MEG-XL in two settings. In the first, we fine-tune MEG-XL on all of the available training data in LibriBrain100; in the second, we exclude the training data of subjects 1–32. Training with the broad subject data appears to be helpful for generalisation to different speech distributions in the form of TIMIT, Mocha-TIMIT, and the podcast data. However, when evaluating on Sherlock, which forms the majority of the data, training without subjects 1-32 led to the best test performance, though only marginally. In general, even when we have a large amount of data for a single subject (Subject 0), additional data from 32 other subjects appears to improve downstream performance on the single subject (Subject 0).

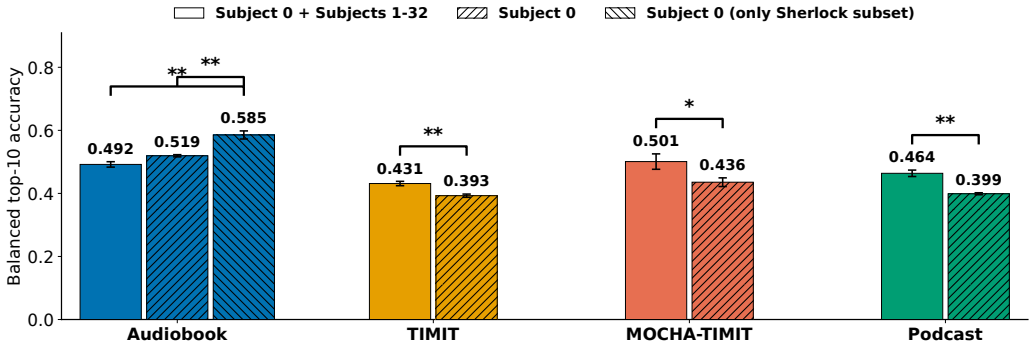

Figure 2: **Subject 0 performance on held-out test data**. Performance improves when training jointly with Subject 1-32 training data, except on the Sherlock subset. Here, training only on Sherlock training data leads to the best performance. Random chance accuracy is 0.2. * indicates $p < .05$ and ** indicates $p < .01$ under Mann-Whitney U-tests.

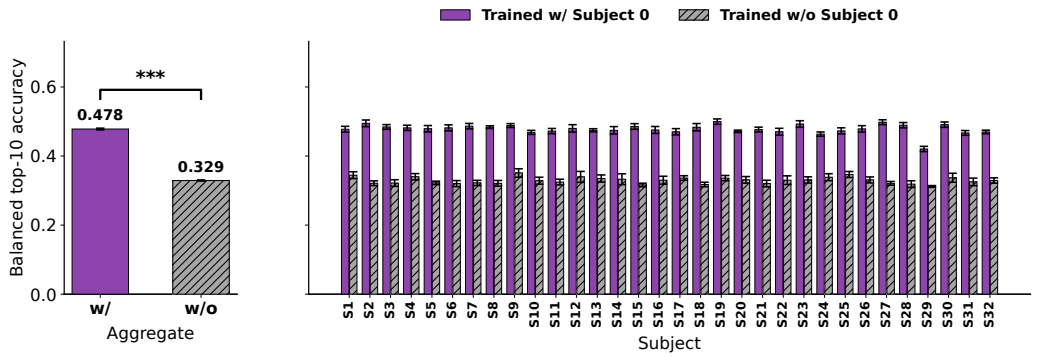

Figure 3: **Fine-tuning MEG-XL on subjects 1-32, with and without training alongside Subject 0**. We report performance on the Sherlock test set. Training jointly with Subject 0 significantly improves generalisation on the broad and shallow data of subjects 1-32. Random chance accuracy is 0.2. *** indicates $p < .001$ under a Mann-Whitney U-test. The per-subject results in the right panel illustrate that difference is not driven by any specific subject, or proper subset of subjects, but is rather consistent across all 32 subjects.

4.2 Task 2: From One-to-Many — Improving Between-Subject Word Classification

In Figure 3, we assess generalisation to the held-out test data on subjects 1–32. Similar to the previous section, we fine-tune MEG-XL in two settings. Here, we analyse training on the multi-subject training data jointly with Subject 0's training data ("Training w/ Subject 0"), and also the effect of excluding the Subject 0 training data ("Training w/o Subject 0"). All subjects generalise similarly, and including Subject 0's training data improves generalisation across subjects by about 15 percentage points. This shows that a large amount of data from a single subject ($\sim$80 hours) can be used to improve performance on many individuals for whom we have $\sim$120$\times$ less data ($\sim$40 minutes each). In the next section, we investigate pushing this discrepancy further, reducing the amount of fine-tuning data needed for subjects 1–32 down to $\sim$10 minutes apiece.

4.3 Task 3: Exporing Efficiency — Maintaining Improvements in Between-Subject Word Classification with Decreasing amounts of Supervised Fine-Tuning Data

Here we investigate the impact of reducing the available fine-tuning data for subjects 1–32, building on the results from Section 4.2. This is an important test for future applications, as even 40 minutes of brain recordings may be too much for patients to use a speech BCI. A speech brain–computer interface should generalise to clinical patients with minimal subject-specific training.

In Figure 4, we compare the use of 100% of the available training data for 12 subjects, with 50% for the next 10 subjects, and 25% for the remaining 10 subjects. In all cases, we see minimal degradation on generalisation, suggesting that cross-subject transfer is responsible for most of the present decoding performance. Our strategy of pre-training on deep single-subject data and fine-tuning to new individuals appears to be sound, and may even be practical for future clinical applications — though we emphasise that all LibriBrain100 data was acquired from healthy volunteers.

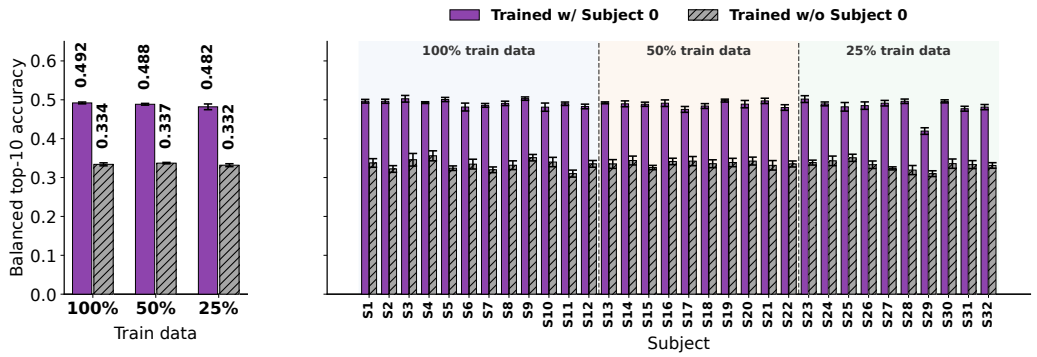

Figure 4: **Performance as a function of fine-tuning data size**. Generalisation remains robust, even with only 25% training data, equivalent to about 10 minutes of brain recordings (i.e., a small enough amount of data to be interesting for clinical applications). The largest difference comes from including data from Subject 0. Random chance accuracy is 0.2. None of the differences in performance under different training data percentages were significant under Mann-Whitney U-tests.

5 Discussion

LibriBrain100 substantially extends the LibriBrain ecosystem in both scale and scope. With $\sim$80 hours from a single subject and $\sim$40 minutes from each of 32 additional subjects, it provides the largest within-subject MEG dataset for speech decoding to date while simultaneously opening the door to cross-subject generalisation research.

Limitations addressed. The original LibriBrain paper closed with a list of limitations. LibriBrain100 directly addresses several of them. First, the *single-subject design*: the addition of 32 new subjects makes cross-subject generalisation a first-class benchmark task for the first time in this dataset series. Second, *focused language content*: whereas LibriBrain drew exclusively from books in the Sherlock Holmes series, LibriBrain100 adds stimuli deliberately spanning the phonetic–semantic axis — TIMIT and MOCHA-TIMIT for controlled phonetic coverage, and podcast narratives for broader semantic diversity (Tang et al., 2023). We expect these sub-datasets will be useful in their own right, enabling targeted studies of phonetic and semantic decoding independently of the collective LibriBrain100 datasets. Third, *preprocessing*: the serialised HDF5 release lowers the barrier to entry for ML researchers by removing the need for domain-specific signal processing knowledge, while the raw BIDS release preserves full flexibility for neuroscientists who wish to apply their own pipelines.

Limitations remaining. Two limitations from the original LibriBrain list remain open. The first is the *listening paradigm focus*. All data in LibriBrain100 were collected during passive listening to continuous speech. Listening provides a natural starting point for speech decoding, as alignments between stimulus and neural response are straightforward to derive, and signal-to-noise is comparatively high. However, it is not the paradigm that ultimately matters for BCIs, where users must attempt to produce or imagine speech. We have been actively collecting inner speech data in parallel and plan to release multiple inner speech datasets going forward. The second remaining limitation is *brain-to-text*. We made a deliberate decision not to include a brain-to-text baseline in this release. Progress on the open-ended decoding problem has been rapid but uneven, and we judged that the field is not yet at a point where a single baseline result would be stable enough to anchor community benchmarking — however, we will revisit this in future releases as methods mature.

Conclusion. LibriBrain100 follows the logic of ImageNet: the value of a shared benchmark compounds over time as more methods are trained and evaluated against it. The open ML competition, public leaderboard, and standardised Python library are all designed with cumulative progress in mind. We hope that LibriBrain100, like its predecessor, will serve not just as a dataset but as shared infrastructure — lowering the cost of entry for ML researchers, enabling rigorous comparison across methods, and ultimately accelerating progress toward practical non-invasive brain-computer interfaces capable of restoring communication to people living with severe paralysis.

Acknowledgements and Disclosure of Funding

We are grateful to our anonymous reviewers for their time and attention, which helped to make this paper better, and to the NeurIPS Evaluations & Datasets organisers for all of their efforts in making the track possible. We acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558), Hartree Centre resources, and the NVIDIA Corporation for donating additional GPUs. PNPL is supported by the MRC (MR/X00757X/1), Royal Society (RG\R1\241267), NSF (2314493), NFRF (NFRFT-2022-00241), SSHRC (895-2023-1022), and ARIA (SCNI-SE01-P004).

References

Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568(7753):493–498.

Armeni, K., Güçlü, U., van Gerven, M., and Schoffelen, J.-M. (2022). A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data, 9:278.

Avramidis, K., Feng, T., Jeong, W., Lee, J., Cui, W., Leahy, R. M., and Narayanan, S. (2025). Neural codecs as biosignal tokenizers. arXiv preprint arXiv:2510.09095.

Bel, C., Bonnaire, J., Pallier, C., and King, J.-R. (2026). "littleprince_meg_french_listen_pallier2025".

Bénar, C.-G., Velmurugan, J., López-Madrona, V. J., Pizzo, F., and Badier, J.-M. (2021). Detection and localization of deep sources in magnetoencephalography: A review. Current Opinion in Biomedical Engineering, 18:100285.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9/10):341–345.

Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., and Brandman, D. M. (2024). An accurate and rapidly calibrating speech neuroprosthesis. New England Journal of Medicine, 391(7):609–618.

Cheung, C., Hamilton, L. S., Johnson, K., and Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. eLife, 5:e12577.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C., and King, J.-R. (2025). Towards decoding individual words from non-invasive brain recordings. Nature Communications, 16:10521.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Doyle, A. C. (1887). A Study in Scarlet. Ward, Lock & Co.

Doyle, A. C. (1890). The Sign of the Four. Spencer Blackett.

Doyle, A. C. (1892). The Adventures of Sherlock Holmes. George Newnes.

Doyle, A. C. (1893). The Memoirs of Sherlock Holmes. George Newnes.

Doyle, A. C. (1902). The Hound of the Baskervilles. George Newnes.

Doyle, A. C. (1905). The Return of Sherlock Holmes. George Newnes.

Doyle, A. C. (1915). The Valley of Fear. George H. Doran Company.

Doyle, A. C. (1917). His Last Bow. John Murray.

Doyle, A. C. (1927). The Case-Book of Sherlock Holmes. John Murray.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence, 5(10):1097–1107.

Elvers, G., Landau, G., Mantegna, F., Özdogan, M., Kim, T., Kwon, T., Cho, S., Ballyk, B., Kurth, L., Jayalath, D., Somaiya, P., de Zuazo, X., Shillingford, B., Farquhar, G., Jiang, M., Jerbi, K., Abdelhedi, H., Mantilla Ramos, Y., Gulcehre, C., Woolrich, M., Voets, N., and Parker Jones, O. (2026). Benchmarking non-invasive speech BCIs: Lessons learned from the 2025 PNPL competition. Journal of Machine Learning Research. In press.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993a). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993b). TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC93S1.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6645–6649. IEEE.

Gwilliams, L., Flick, G., Marantz, A., Pylkkanen, L., Poeppel, D., and King, J.-R. (2023). Introducing MEG-MASC: A high-quality magneto-encephalography dataset for evaluating natural speech processing. Scientific Data, 10(1):862.

Gwilliams, L., King, J.-R., Marantz, A., and Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. Nature Communications, 13(1):6606.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of Modern Physics, 65(2):413–497.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97.

Jayalath, D., Ballyk, B., and Parker Jones, O. (2026). On the problem of measuring progress in speech brain-computer interfaces. arXiv preprint. To appear.

Jayalath, D., Landau, G., and Parker Jones, O. (2025a). Unlocking non-invasive brain-to-text. arXiv preprint arXiv:2505.13446.

Jayalath, D., Landau, G., Shillingford, B., Woolrich, M. W., and Parker Jones, O. (2025b). The Brain's Bitter Lesson: Scaling speech decoding with self-supervised learning. In Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025, volume 267 of Proceedings of Machine Learning Research. PMLR.

Jayalath, D. and Parker Jones, O. (2026). MEG-XL: Data-efficient brain-to-text via long-context pre-training. In Forty-third International Conference on Machine Learning, ICML 2026, Seoul, South Korea, July 6-11, 2026, Proceedings of Machine Learning Research. PMLR.

Jiang, W., Zhao, L., and Lu, B. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI. In International Conference on Learning Representations (ICLR).

King, J.-R., Bel, C., Evanson, L., Gadonneix, J., Houhamdi, S., Lévy, J., Raugel, J., Santos Revilla, A., Zhang, M., Bonnaire, J., Caucheteux, C., Défossez, A., Desbordes, T., Diego-Simón, P., Khanna, S., Millet, J., Orhan, P., Panchavati, S., Ratouchniak, A., Thual, A., Brooks, T. L., Begany, K., Benchetrit, Y., Careil, M., Banville, H., d'Ascoli, S., Dahan, S., and Rapin, J. (2026). NeuralSet: A high-performing Python package for Neuro-AI. Preprint; URL will be updated when the paper lands on arXiv.

Landau, G., Özdogan, M., Elvers, G., Mantegna, F., Somaiya, P., Jayalath, D., Kurth, L., Kwon, T., Shillingford, B., Farquhar, G., Jiang, M., Jerbi, K., Abdelhedi, H., Mantilla Ramos, Y., Gulcehre, C., Woolrich, M., Voets, N., and Parker Jones, O. (2025). The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. NeurIPS, Competition Track. https://arxiv.org/abs/2506.10165.

Makin, J. G., Moses, D. A., and Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder-decoder framework. Nature Neuroscience, 23(4):575–582.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174):1006–1010.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., and Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620:1037–1046.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Mohamed, A.-r., Dahl, G. E., and Hinton, G. (2009). Deep belief networks for phone recognition. In NIPS Workshop on Deep Learning for Speech Recognition and Related Applications.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., and Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3):217–227.

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., Henson, R. N., Jas, M., Litvak, V., Moreau, J. T., Oostenveld, R., Schoffelen, J.-M., Tadel, F., Wexler, J., and Baillet, S. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. Scientific Data, 5:180110.

Ochshorn, R. M. and Hawkins, M. (2015). Gentle: A robust yet lenient forced aligner built on Kaldi.

Peelle, J. E. and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. Frontiers in Psychology, 3:320.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2):8–13.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US. IEEE Signal Processing Society.

Ryle, G. (1949). The Concept of Mind. Hutchinson, London.

Schoffelen, J.-M., Oostenveld, R., Lam, N. H. L., Uddén, J., Hultén, A., and Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. Scientific Data, 6:1–13.

Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J. S., Dalgleish, T., Henson, R. N. A., Brayne, C., and Matthews, F. E. (2014). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurology, 14.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience, 26(5):858–866.

Taulu, S. and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. Physics in Medicine & Biology, 51(7):1759.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., and Li, H. (2024). EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In The Thirty-Eighth Annual Conference on Neural Information Processing Systems.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. (2025). CBraMod: A criss-cross brain foundation model for EEG decoding. International Conference on Learning Representations (ICLR).

Wang, S., Zhang, X., Zhang, J., and Zong, C. (2022). A synchronized multimodal neuroimaging dataset for studying brain language processing. Scientific Data, 9.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., and Henderson, J. M. (2023). A high-performance speech neuroprosthesis. Nature, 620:1031–1036.

Wrench, A. (1999). The MOCHA-TIMIT articulatory database. https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html. Created November 1999; accessed 2026-02-07.

Wrench, A. A. (2000). A multi-channel/multi-speaker articulatory database for continuous speech recognition research. Phonus, 5:1–13.

Xiao, Q., Cui, Z., Zhang, C., Chen, S., Wu, W., Thwaites, A., Woolgar, A., Zhou, B., and Zhang, C. (2025). BrainOmni: A brain foundation model for unified EEG and MEG signals. In The Thirty-Ninth Annual Conference on Neural Information Processing Systems.

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press, Cambridge, MA.

Özdogan, M., Landau, G., Elvers, G., Jayalath, D., Somaiya, P., Mantegna, F., Woolrich, M., and Parker Jones, O. (2025). LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. NeurIPS, Datasets and Benchmarks Track.

Appendices / Supplemental Materials

A Additional Dataset Details

A.1 Visual Summary

Figure 5 provides two complementary views of the LibriBrain100 dataset composition. Panel (a) shows recording hours by subject and corpus; panel (b) shows a treemap of the full dataset with rectangle area proportional to duration. See Table 1 for a complementary perspective.

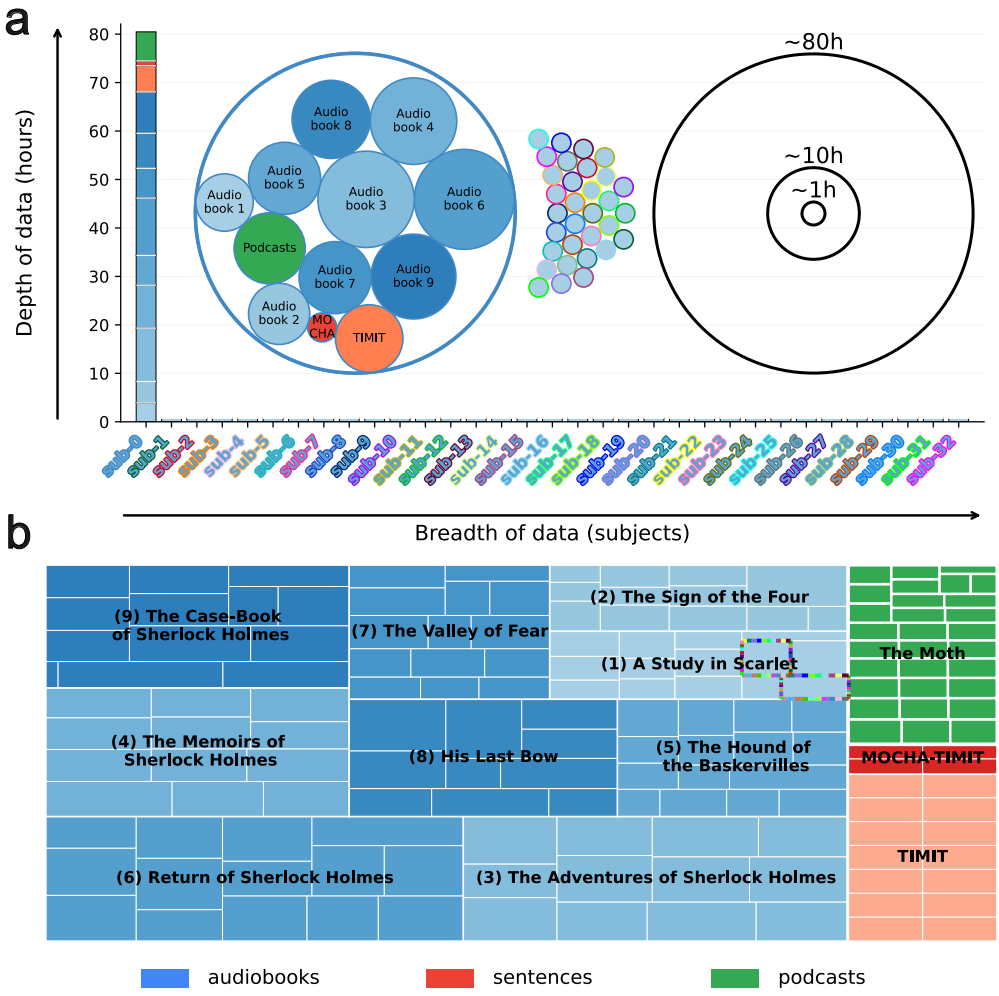

Figure 5: **Dataset proportions**.(**a**) Recording hours by subject and corpus. The stacked bar plot shows total duration per subject, with sub-0 comprising multiple linguistic materials and other subjects a subset of audiobook data. The bubble plot represents the same information (area proportional to hours); the larger sub-0 circle is subdivided into audiobooks, sentences, and podcasts, while smaller circles represent other subjects. Colours indicate data type, blue shades distinguish between audiobooks, and circle outlines distinguish between subjects. Black concentric circles provide a size reference (~1 h, ~10 h, ~80 h). (**b**) Treemap of the full dataset, with rectangle area proportional to duration. Categories are subdivided into corpora and then into MEG recording sessions. The multicolour outline marks the only two audiobook sessions recorded from multiple subjects.

A.2 Comparison with Existing Datasets

Table 2 lists existing MEG datasets for speech comprehension at the time of writing. LibriBrain100 compares favourably in total hours (104) and hours per subject (0.6–80), with the highest scores (by far, in the case of maximum hours per subject). It sits comfortably in the middle of the pack in terms of number of subject, with 33 (range 1–96). The `pnpl` library contains data loaders for all of these datasets with support for reproducible data splits and for dataset downloading where possible, to support machine learning at scale.

Table 2: **Breadth and depth of speech comprehension MEG datasets**.

Name	Dataset	Language	Sensors	Total hrs	# Subj	Hrs/Subj	Stimulus	Public	PNPL Dataloader
MOUS	Schoffelen et al. (2019)	Dutch	275	81	**96**	0.8	Spoken sentences	Yes	Yes
MEG-MASC	Gwilliams et al. (2022)	English	208	49	27	1.0	MASC stories (synthetic voice)	Yes	Yes
Armeni	Armeni et al. (2022)	English	269	30	3	10.0	LibriVox (Sherlock Book 3)	Yes	Yes
Le Petit Prince	Bel et al. (2026)	French	**306**	93	58	1.6	Le Petit Prince	Yes	Yes
LibriBrain	Özdogan et al. (2025)	English	**306**	52	1	52.3	LibriVox (Sherlock Book 1–7)	Yes	Yes
LibriBrain100	(ours)	English	**306**	**104**	33	0.6–**80.0**	LibriVox (Sherlock Book 1–9), TIMIT, MOCHA-TIMIT, The Moth (30 Podcasts)	Yes	Yes

B Data Collection Methods

B.1 Subjects

MEG recordings were acquired from 33 volunteers (10 female, 23 male; age range 19–50 years, median = 28, interquartile range = 8). All participants reported normal hearing and normal or corrected-to-normal vision, and none had a history of neurological disorders. Sixteen participants were native speakers of English, while the remaining seventeen acquired English as a second language but were highly proficient and reported daily use for work or study. Prior to participation, all individuals provided informed consent for the use of anonymised data for research purposes. The study was approved by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (R90053/RE002).

B.2 Stimulus Materials

To prepare the audio stimuli for the MEG experiments, all audio files were converted to uncompressed WAV format and resampled to 48 kHz using SoX. The goal was to segment the audio into natural sentences or phrases, typically defined by pauses, and to align each segment with its corresponding textual content. This enabled the delivery of precise triggers to the MEG system during stimulus presentation, allowing continuous verification of temporal alignment between MEG recordings, audio signals, and linguistic annotations. Audio segmentation was carried out in two stages: an initial automated phase followed by manual refinement. The automated segmentation relied on voice activity detection (VAD), while the manual stage corrected segment boundaries in cases where VAD failed, for example by truncating low-intensity sounds such as utterance-initial voiceless plosives. In addition, manual refinement involved checking and reconciling consistency between data-driven audio segmentation and text-based boundaries (e.g., punctuation and sentence structure). VAD was implemented using custom scripts in Praat (Boersma, 2001), identifying speech segments based on an intensity threshold of 59 dB and a minimum duration of 600 ms. The resulting TextGrid annotations were then populated with the corresponding text and carefully reviewed and corrected by hand. This process required over 200 hours of expert human effort, exceeding standard practices in the design and curation of comparable datasets, and was undertaken to ensure precise alignment across MEG recordings, audio signals, and linguistic annotations.

Audiobooks

Audio recordings for all nine books in the canon of Sherlock Holmes were sourced from LibriVox (https://librivox.org/). The books (including both novels and short stories) were presented in chronological order of publication. To minimise variance, we utilised recordings from the same reader (David Clarke) for books 1–8. As audio recordings were not available from this speaker for book 9, we selected a second reader with similar characteristics (Thomas A. Copeland; British English male).

Each MEG recording session corresponded to a standalone chapter from the audiobooks. All texts are in the public domain and were obtained from Project Gutenberg (https://www.gutenberg.org/). Manual correction was used to fix typos and, where text and audio diverged, to align the text with the spoken audio. Ambiguous text, such as numbers, was normalised (e.g., "1492" rendered as "fourteen ninety-two" if spoken that way). Details for the audiobooks are provided in Table 3.

The Sherlock Holmes audiobooks comprise multiple detective fiction stories focused on solving mysteries through observation and logical deduction. Narratives typically involve identifying and interpreting clues, with cases ranging from concrete, crime-based investigations to stories that initially present more unusual or seemingly supernatural elements before being resolved through reasoning. The prose includes a mixture of narration and both direct and indirect dialogue. The thematics are relatively homogeneous (e.g., crime, investigation), but variation arises across books and chapters through differences in setting, narrative context, and secondary characters. The recordings are read by professional narrators, resulting in a controlled delivery: pauses tend to align with punctuation and sentence structure, diction is clear and consistent, and breathing and other non-linguistic artefacts are minimised.

Table 3: **Audiobooks available in LibriBrain100**. Books 1–7 appeared in the original LibriBrain release (Özdogan et al., 2025); the rest are new here. Text in the name column link to the source audio on LibriVox.

Book	Name	Type	Sessions	Hours
1	A Study in Scarlet (1887)	Novel	14	04:37:34
2	The Sign of the Four (1890)	Novel	12	04:27:31
3	The Adventures of Sherlock Holmes (1892)	Short Stories	12	10:56:13
4	The Memoirs of Sherlock Holmes (1894)	Short Stories	12	08:53:17
5	The Hound of the Baskervilles (1901–1902)	Novel	15	06:10:32
6	Return of Sherlock Holmes (1905)	Short Stories	14	11:51:17
7	The Valley of Fear (1914–1915)	Novel	14	06:06:17
8	His Last Bow (1917)	Short Stories	10	07:10:29
9	The Case-Book of Sherlock Holmes* (1927)	Short Stories	13	08:27:01
				68:40:11

*Read by Thomas A. Copeland

TIMIT

TIMIT is a corpus of read American English speech designed for acoustic–phonetic research and the development and evaluation of automatic speech recognition systems (Garofolo et al., 1993a,b). It contains recordings from 630 speakers across 8 major dialect regions of the United States, each producing 10 sentences, for a total of 6300 utterances (∼5 hours of speech). All speakers were native speakers of American English and were screened to exclude clinically significant speech pathology. Speaker recruitment aimed to cover major dialect regions based on the region in which speakers lived during childhood. While the corpus achieves broad dialectal coverage, it is not perfectly balanced: some regions (e.g., New England, New York City, and "Army Brat") are underrepresented due to practical constraints. The corpus is also sex-imbalanced (438 male, 192 female; approximately 70%/30%).

The sentence prompts were designed to provide controlled phonetic coverage. Each speaker reads 2 shared SA ("dialect") sentences, 5 SX sentences selected from a set of 450 phonetically compact sentences (each repeated across 7 speakers), and 3 SI sentences drawn from a set of 1890 phonetically diverse sentences (each produced by a single speaker), yielding 2342 distinct sentence texts in total. The SX and SI sentences were constructed to maximise coverage of phonetic contexts and allophonic variation, rather than semantic diversity. Speech was recorded at 16 kHz using 16-bit linear PCM encoding. In this dataset release, we provide MEG recordings for the full set of 6300 utterances.

The TIMIT corpus consists of short, isolated sentences with no shared discourse context across utterances. The recordings are read speech, produced in a controlled setting, with clear articulation, limited disfluencies, and pauses aligned with sentence boundaries. At the same time, variability is introduced through the large number of speakers and dialect regions, providing diversity in pronunciation, accent, and voice characteristics. This combination of tightly controlled phonetic design and

speaker variability makes TIMIT particularly well-suited for analyses focused on acoustic–phonetic representations.

MOCHA-TIMIT

MOCHA-TIMIT is a multichannel articulatory–acoustic corpus designed to support research on speech production and acoustic–articulatory modelling (Wrench, 1999, 2000). The publicly distributed release contains data from two speakers: one male (msak0) and one female (fsew0), both Southern British English speakers. In this dataset release, we provide all 460 utterances from each speaker (920 utterances total). The sentence set was constructed to provide broad phonetic coverage while explicitly targeting connected-speech processes in English, such as assimilation and weak forms, and was designed to parallel the phonetic coverage of TIMIT while reflecting British English pronunciation. Although the MOCHA-TIMIT corpus was originally developed for studying speech production, the tight control over articulation and phonetic content also makes it well-suited for perception studies, as it provides highly structured and phonetically balanced stimuli with well-characterised acoustic realisations.

Each utterance is a short, read sentence produced in a controlled recording environment. As in TIMIT, there is no extended narrative context across sentences; instead, the corpus emphasises phonetic and articulatory diversity within isolated utterances. The recordings are carefully produced, with clear diction, minimal disfluencies, and pauses aligned with sentence boundaries. Compared to TIMIT, however, the speech more systematically reflects connected-speech phenomena, resulting in more natural coarticulation patterns despite the controlled setting. For example, assimilation processes may occur across word boundaries (e.g., "good boy" realised with a more bilabial /d/ influenced by the following /b/), and weak forms are frequently used for function words (e.g., "and" $\rightarrow$ /ən/, "to" $\rightarrow$ /tə/, "of" $\rightarrow$ /əv/). Vowel reduction and consonant lenition are also present in unstressed positions, and segmental realisations are influenced by surrounding phonetic context, yielding more continuous and context-dependent articulatory patterns.

Podcasts

We used 30 stories (6 hours) from The Moth podcast, drawn from a larger set of 77 stories ($\sim$15 hours) previously used for semantic decoding from fMRI data (Tang et al., 2023). This subset was selected using a data-driven criterion aimed at maximising semantic coverage. Specifically, we estimated the spread of each story in semantic space using sentence embedding models, which map sentences to dense vector representations that capture their semantic content. For each story, we computed the centroid of its sentence embeddings and quantified dispersion as the mean cosine distance of individual sentence vectors to this centroid in the high-dimensional embedding space. Stories with greater dispersion (i.e., covering a wider range of semantic content, as opposed to being concentrated in a narrow region of the embedding space) were preferentially selected. Details for the selected stories are provided in Table 4.

The Moth podcast stories provide an effective set of naturalistic stimuli for neural decoding, in particular for semantic representations, because they combine ecological validity with high semantic diversity and coherent narrative structure. Each stimulus consists of a single speaker telling an autobiographical narrative, ensuring continuity in voice, perspective, and discourse structure. Crucially, the corpus spans a wide range of topics and contexts, from historically and geographically specific settings (e.g., Zimbabwe during independence, wartime Baghdad, travel in the USSR, scientific expeditions) to personal and emotional experiences (e.g., bereavement, parenthood, moral dilemmas), as well as humorous, reflective, and self-development narratives. This breadth yields rich variation in entities, events, and abstract themes, providing dense coverage of semantic space. At the same time, individual stories are internally coherent and contextually well-defined, allowing models to track how meaning unfolds over extended timescales.

These narratives are delivered as spontaneous speech rather than read text, and therefore include natural disfluencies such as interjections (e.g., "like," "you know," "I mean"), repetitions (e.g., "and and then…," "I was I was…"), and repairs (e.g., "on Tuesday—uh, Wednesday," "her brother—no, her cousin"), which are largely absent in scripted speech. Temporal dynamics also differ from read speech: pauses can reflect on-the-fly planning, while other segments are produced in dense, continuous stretches, a characteristic feature of spontaneous narration. In addition, prosodic variation is more pronounced, as speakers may raise their voice or modulate intonation for emphasis or comedic effect.

Recordings also include audience responses such as laughter and applause, which introduce acoustic variability but can carry contextual and pragmatic information tied to the narrative. Together, these properties enhance ecological validity and introduce variability that better reflects real-world language processing. The combination of rich thematic diversity, extended context, and naturalistic delivery, as leveraged in prior work (e.g., Tang et al., 2023), makes Moth stories particularly well-suited for training and evaluating neural decoders that map brain activity to continuous, context-sensitive semantic representations.

Table 4: **Podcasts available in LibriBrain100**. Story titles are hyperlinked to the source audio and transcript at https://themoth.org/stories/. Validation and test set stories are highlighted yellow and red, respectively. Durations are given in minutes and seconds for the audio files used in our stimulus set.

#	Story	Author	Duration
1	Alternate Ithaca Tom	Tom Weiser	11:47
2	Thumbs Up!	Nathan Englander	14:09
3	Goldie, the Goldfish	Becca Stevens	10:54
4	Under the Influence	Jeffery Rudell	10:28
5	Treasure Island	Boots Riley	13:30
6	That Thing on My Arm	Padma Lakshmi	14:49
7	Breaking Up in the Age of Google	Jessi Klein	17:43
8	The Triangle Shirtwaist Connection	Michelle Fecteau	7:06
9	Stage Fright	Suzanne Vega	10:07
10	Not on the Usual Tour	Jon Lovett	8:31
11	Life Reimagined	Raymond Christian	11:15
12	Only One Way To Find Out	Sarah Gray	13:04
13	The Postman Always Calls	Elizabeth Browning	15:29
14	Blue Hope	Sylvia Earle	14:00
15	The Curse	Dame Wilburn	13:56
16	Beneath the Mushroom Cloud	Clifton Truman Daniel	11:46
17	Going the Liberty Way	Kevin Roose	13:28
18	Caution: Eating	Evan Kleiman	9:40
19	Birth of a Nation	Petina Gappah	9:09
20	Life and Death on the Oregon Trail	Micaela Blei	12:24
21	Fire Test For Love	Lauren Slater	10:58
22	Leaving Baghdad	Abbas Mousa	11:15
23	Coming of Age on Death Row	Gautam Narula	11:58
24	How To Draw A Nekkid Man	Tricia Rose Burt	12:08
25	The Mayor of the Freaks	David Crabb	16:11
26	Swimming with Astronauts	Michael J. Massimino	13:11
27	Wild Womxn and Dancing Queens	Lex Jade	6:40
28	Vixen and the USSR	Sue Steinacher	13:24
29	From Boyhood to Fatherhood	Jonathan Ames	11:57
30	Where There's Smoke	Jenifer Hixson	10:02
			6:00:59

B.3 Experimental Design & Procedure

Each recording session began with visually presented instructions projected onto a translucent whiteboard using a DLP LED projector (ProPixx, VPixx Technologies Inc., Saint-Bruno, Canada). Participants were seated inside the MEG scanner and initiated the experiment via button press. Auditory stimuli were delivered binaurally through non-metallic air tube earphones (Aero Technologies) at approximately 70 dB SPL, with minor adjustments to bass and treble based on participant preference. Stimulus delivery was controlled using the PsychoPy toolbox (Peirce, 2007), and the stimulus computer synchronized with the MEG acquisition system via a parallel port to send precise event triggers marking stimulus onset with millisecond temporal accuracy.

For the multiple subjects cohort (sub-01 to sub-32), all participants listened to Chapters 11 and 12 of A Study in Scarlet. Both chapters were presented within a single recording session lasting approximately one hour, with a short break between chapters. To assess attention and comprehension, participants answered five questions at the end of each chapter. Each question was presented in a four-alternative multiple-choice format with distractors (e.g., "What does Jefferson Hope take from Lucy Ferrier after her death?" with answer options "A: A locket, B: A necklace, C: A wedding ring, D: A bracelet"). Responses were recorded via button press using a MEG-compatible ResponsePixx Dual Handheld system (VPixx Technologies Inc., Saint-Bruno, Canada).

Subject 0 participated in multiple recording sessions and was exposed to a broader set of linguistic materials, including all books in the Sherlock Holmes canon, as well as additional speech corpora (e.g., TIMIT, MOCHA-TIMIT, and The Moth podcasts). Comprehension assessment varied by stimulus type. For audiobook stimuli, a single comprehension question was presented at the end of each chapter

in a two alternative options format (e.g., "Where is the body of the murder victim found?" with answer options "A: in the bedroom", "B: in the garden"). For TIMIT and MOCHA-TIMIT, questions were presented immediately after selected sentences (approximately 20 questions per session; 5% of trials for TIMIT, 10% for MOCHA-TIMIT). These questions were presented in a four alternative multiple options format (e.g., "_____ is a pop singer." with answer options "A: Michael Jackson, B: Tina Turner, C: Elton John, D: Madonna"; "Clear _____ is appreciated." with answer options "A: grammar, B: diction, C: articulation, D: pronunciation") and typically targeted a key word (often a noun), with distractors that were either acoustically or semantically similar. For the podcast corpus, five comprehension questions were presented at the end of each episode, probing key details, event sequences, and overall meaning (e.g., "How did thinking about his alternate life affect him?" with answer options "A: It made him confident, B: It made him confused and unhappy, C: It motivated him to study, D: It made him excited"). For subject 0, each recording session lasted approximately three hours and typically included either 3–5 audiobook chapters, 1000–1200 sentences, or 8–10 podcast episodes. Short breaks were allowed between recording blocks. Recording sessions were spaced at least one day apart, with no more than two months between sessions, depending on participant and experimenter availability.

B.4 Data Acquisition

Prior to data acquisition, each participant's head shape was digitised using a Polhemus Fastrak 3D digitiser (Polhemus, Vermont, USA). This procedure included the localisation of fiducial landmarks (nasion, left and right pre-auricular points), along with approximately 300 additional points sampled across the scalp, forehead, and nose. Five Head Position Indicator (HPI) coils were positioned on the mastoid and forehead to enable continuous monitoring of head position during MEG recording via electromagnetic induction. MEG recordings were acquired using a MEGIN Triux™ Neo system (York Instruments Ltd., Heslington, UK), consisting of 102 magnetometers and 204 orthogonal planar gradiometers. The system was installed in a magnetically shielded room to reduce environmental interference. Prior to entering the recording environment, participants were screened for metallic objects and other potential sources of electromagnetic noise. During acquisition, participants were seated with their head positioned close to the dewar. Participants were instructed to minimise head, body, and limb movements throughout the recording. Data were sampled at 1000 Hz with an online band-pass filter of 0.01–330 Hz. Ocular activity was monitored using bipolar electrooculogram (EOG) electrodes, with one pair placed at the outer canthi (horizontal EOG) and another above and below the left eye (vertical EOG). Cardiac signals were recorded via bipolar electrocardiogram (ECG) electrodes positioned on the clavicle and hip. Articulatory muscle activity was continuously tracked using electromyography (EMG), with electrodes placed below the cheekbone (jaw movement), below the lower lip (lip movement), and beneath the chin to capture potential laryngeal activity.

B.5 Minimal Preprocessing Pipeline

The MEG data were minimally preprocessed to remove measurement artefacts while preserving flexibility for additional preprocessing steps for downstream analyses. Head position was estimated using continuous recordings from the Head Position Indicator (HPI) coils throughout each recording session. This information was used to correct for head movements, and data from all participants were realigned to a common reference head position. This procedure was applied consistently across subjects and is expected to reduce variability due to head position differences and within-session head movement. Noisy channels were identified, excluded, and subsequently reconstructed via interpolation from neighbouring sensors. Environmental noise was attenuated using Maxwell filtering. Specifically, a temporally non-extended Signal Space Separation (SSS) algorithm (Taulu and Simola, 2006) was applied to suppress sources originating outside the head. In contrast, no explicit steps were taken to remove physiological artefacts (e.g., eye blinks, cardiac activity, or muscle contractions), allowing users to apply additional artefact correction procedures as appropriate for their specific downstream analyses. Power line noise at 50 Hz and its 100 Hz harmonic was removed using notch filters. The data were then band-pass filtered between 0.1 and 125 Hz using zero-phase, two-pass Butterworth filters to reduce slow drifts and prevent aliasing effects prior to downsampling. Finally, the data were downsampled to 250 Hz, yielding recordings with 306 channels and a temporal resolution of 4 ms per sample.

B.6 Event Files/Labels

Annotations were generated for each session, providing onset times and durations (in seconds) for multiple event types, including silence (i.e., non-speech segments, which may include breathing and occasional laughter or applause), word-level units (e.g., A, Study, in, Scarlet), and phoneme-level units (e.g., ah, s, t, ah, d, iy, ih, n, s, k, aa, r, l, ah, t). Additional annotation fields include phoneme position within words, encoded using conventions from Kaldi (Povey et al., 2011): B (beginning), I (inside), E (end), and S (singleton). Each MEG recording session (available as a FIF file) is accompanied by a corresponding event file in TSV format, which can be used to define labels for supervised learning tasks. These event files were derived from the transcripts and their corresponding audio segments.

To obtain precise temporal alignment between text and audio, forced alignment was performed using Gentle (Ochshorn and Hawkins, 2015), a Kaldi-based aligner that employs Gaussian Mixture Model–Hidden Markov Model (GMM-HMM) architectures to combine acoustic and language models and determine the most likely alignment path between audio and transcript. Because the audio had already been segmented into short utterances using VAD, as described above, the forced alignment step was simplified. Instead of processing entire chapters, the aligner was applied to short, pre-segmented audio–transcript pairs, improving both robustness and alignment accuracy.

Gentle produces phoneme-level annotations in the ARPABET format, consisting of a set of 39 phoneme categories. ARPABET provides a discrete, standardized phonemic representation that captures the more abstract, relatively invariant characteristics of speech sounds, but does not capture finer phonetic detail such as allophonic variation or diphthong structure. The same forced-alignment procedure was applied consistently across all linguistic materials (audiobooks, TIMIT, MOCHA-TIMIT, and podcasts) to ensure internal consistency within the dataset. Although alternative phonetic annotations (e.g., IPA transcriptions) are available in the standard releases of the TIMIT and MOCHA-TIMIT corpora, they were not used to generate event files in this dataset. Instead, a unified ARPABET representation was adopted to maintain consistency across heterogeneous linguistic materials.

The forced aligner occasionally produced suboptimal outputs, particularly in challenging cases. Failures were most frequent for proper names and other out-of-vocabulary (OOV) items, quotations in different languages, and segments with atypical prosody (e.g., voice imitation or exaggerated emphasis). Additional difficulties arose in more spontaneous speech (e.g., podcasts), including repetitions, repairs, and unclear or reduced pronunciations. These cases were identified and corrected through manual inspection. Audio segments were reviewed by human experts using Praat (Boersma, 2001), with spectrograms and waveforms examined to refine word and phoneme boundaries and recover missed lexical items. As a result of this intervention, all word-level annotations were recovered. Following manual correction, the proportion of OOV annotations was substantially reduced for audiobooks and podcasts. Overall, this process resulted in high-quality, densely annotated data suitable for both word-level and phoneme-level analyses.

B.7 Standard Data Splits

Audiobooks. For Subject 0, we reuse the same validation and test splits defined in LibriBrain (Özdogan et al., 2025). Here, Sherlock book 1, session 11 is used for validation and session 12 is used for test. As subjects 1-32 have no official training data, when training on these subjects, we use the first half of book 1, session 11 for training, the second half for validation, and keep session 12 independent for testing. When training multi-subject data jointly with Subject 0, we also use only the second half of Subject 0's session 11 for validation, adding the first half to the training set.

TIMIT. We use the official TIMIT core test speaker set of 24 speakers for our test set and 50 speakers from the dev split (Garofolo et al., 1993b). We exclude all `SA` labelled sentence IDs from testing and validation as these are repeated by all speakers (including those in training). We also exclude any speakers from training who have sentences that overlap with the speakers in the test set. See Table 5 for a complete list of the speakers we use in validation and test.

MOCHA-TIMIT. As there is no standard split for MOCHA-TIMIT, we define our own split, being careful to avoid any sentence stimulus overlap while maintaining an independent test session. We note that MOCHA-TIMIT contains four different sets: A, B, C, and D where sentences are repeated between A and D, and between B and C in shuffled orders. We use all of set A and D for training,

Table 5: **TIMIT speaker IDs used for the test and validation splits**.

Test set (24 speakers)				
mdab0	mwbt0	felc0	mtas1	mwew0
fpas0	mjmp0	mlnt0	fpkt0	mlll0
mtls0	fjlm0	mbpm0	mklt0	fnlp0
mcmj0	mjdh0	fmgd0	mgrt0	mnjm0
fdhc0	mjln0	mpam0	fmld0	
Validation set (50 speakers, from dev split)				
faks0	fdac1	fjem0	mgwt0	mjar0
mmdb1	mmdm2	mpdf0	fcmh0	fkms0
mbdg0	mbwm0	mcsh0	fadg0	fdms0
fedw0	mgjf0	mglb0	mrtk0	mtaa0
mtdt0	mthc0	mwjg0	fnmr0	frew0
fsem0	mbns0	mmjr0	mdls0	mdlf0
mdvc0	mers0	fmah0	fdrw0	mrcs0
mrjm4	fcal1	mmwh0	fjsj0	majc0
mjsw0	mreb0	fgjd0	fjmg0	mroa0
mteb0	mjfc0	mrjr0	fmml0	mrws1

then split B and C into validation and test as follows. We take the first half of the sentences in set B as the validation set. We then use the sentences in set C that are not in the first half of set B as test data, noting that these test sentences will not be consecutive within set C. In the Hugging Face repository, sets A, B, C, D correspond to sessions 1, 2, 3, 4.

Podcasts. Following Tang et al. (2023), two stories were designated to be held-out from training. For reproducibility we thus included *From Boyhood to Fatherhood* in our validation set and *Where There's Smoke* in our test set. All other stories should be included in training.

C Additional Stimulus Details

C.1 Linguistic Variability

One of the central objectives of this dataset is to enable rigorous tests of generalization in neural decoding of speech perception. In the context of MEG, this is particularly important because models are exposed to substantial variability in natural speech, spanning differences in speakers, phonetic realisations, and semantic content. Rather than aiming to isolate invariant representations, the goal is to evaluate how well models can achieve robust performance in the presence of such variability, potentially leveraging both variable and more stable features of the signal. To this end, the dataset was designed to span multiple dimensions of variability that are known to shape speech processing, including speaker-dependent acoustic properties, phonetic composition, and higher-level semantic structure. By systematically varying these factors across complementary corpora, we can assess the extent to which decoding models remain resilient across changes in linguistic context and input statistics. This is especially relevant for tasks such as phoneme and word classification, where strong performance should be maintained despite differences in speakers, phonetic distributions, and semantic environments. In addition, the combination of variability and scale allows us to examine how exposure to diverse speech inputs influences learning dynamics, for instance whether it supports faster convergence toward more robust representations that better reflect the range of conditions encountered in natural speech perception.

At the acoustic level, one of the main goals of the dataset is to introduce substantial variability in speaker characteristics, thereby enabling a more realistic assessment of robustness in speech perception. To quantify this variability, we performed a speaker embedding analysis using a pretrained ECAPA-TDNN model (Desplanques et al., 2020). These embeddings capture voice-related acoustic properties such as pitch, timbre, and vocal tract characteristics, providing a compact representation of speaker differences. When projected into a low-dimensional space (Figure 6), the embeddings reveal a clear separation between male and female speakers, indicating that they capture

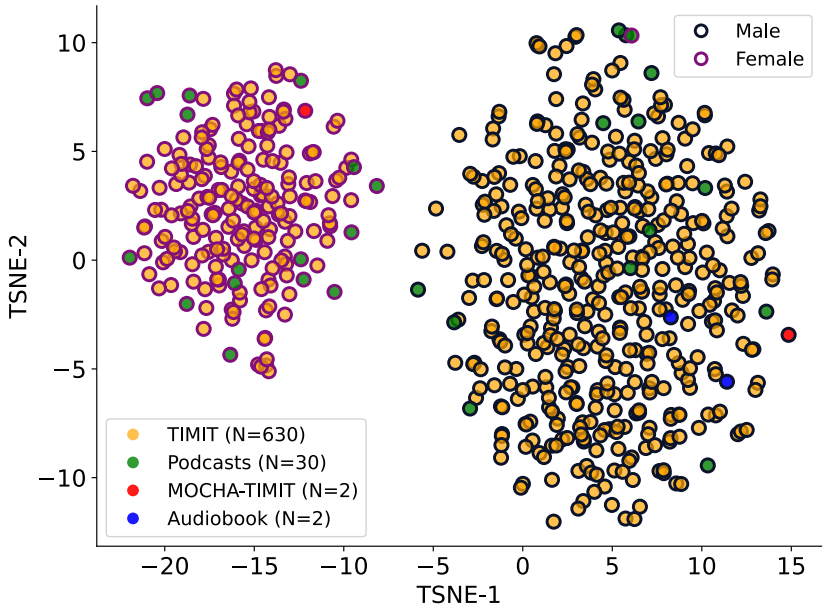

Figure 6: **Speaker embeddings**. Two-dimensional t-SNE projection of ECAPA-TDNN speaker embeddings, which capture acoustic properties such as timbre, pitch, and vocal tract characteristics. Each point represents a single speaker, obtained by averaging embeddings across multiple utterances. Marker fill colour indicates the speech corpus (TIMIT, Podcasts, MOCHA-TIMIT, Audiobook), while marker edge colour denotes speaker sex (male, female).

meaningful acoustic structure. This separation is expected, as one of the most salient differences in speech arises from fundamental frequency (F0) and vocal tract size: on average, male speakers have larger vocal tracts and produce lower-pitched sounds, whereas female speakers tend to have smaller vocal tracts and higher-pitched voices. Beyond this dominant axis, the embeddings also reflect finer-grained differences between individual speakers, consistent with their intended role in capturing speaker identity. Notably, TIMIT spans a broader region of this space compared to the other corpora, indicating wider coverage of speaker variability, in part due to its large number of speakers. Combining all corpora further expands this coverage, resulting in a richer sampling of pitch- and timbre-related dimensions. This is particularly relevant for MEG neural decoding, as such acoustic features are represented in auditory cortex and contribute to the variability of the neural signal. By exposing models to this range of speaker-dependent variation, the dataset allows us to assess how well decoding performance is maintained across acoustic differences, and establishes a foundation for variability not only at the acoustic level but also in the phonetic realisations of speech.

At the phonetic level, this variability extends to the distribution and realisation of speech sounds, providing a complementary axis along which robustness can be evaluated. TIMIT, podcasts, and audiobooks offer distinct phonetic properties, with TIMIT in particular engineered to achieve a more balanced coverage of phonemes, including relatively rare sounds. This is accomplished by repeatedly including sentences that contain less frequent phonemes, thereby increasing their occurrence while preserving the structure of natural speech. As shown in Fig. 7, using approximately 200,000 phoneme tokens per corpus, noticeable differences emerge in the frequency of rare phonemes such as G, Y, SH, OY, and ZH compared to the other corpora. This pattern is also reflected in the rank–frequency analysis (Fig. 8), where the fitted slopes indicate a flatter distribution for TIMIT (-0.7271) relative to Podcasts (-0.7796) and Audiobook (-0.8254), implying reduced dominance of high-frequency phonemes and increased representation of low-frequency ones. Crucially, the large number of speakers in TIMIT further amplifies phonetic variability by introducing a wide range of accents, speaking styles, and speaker-specific articulations. As a result, individual phonemes are realised through a richer set of allophonic variants, and co-articulatory patterns vary more extensively across contexts. This combination of balanced phoneme frequencies and diverse phonetic realisations

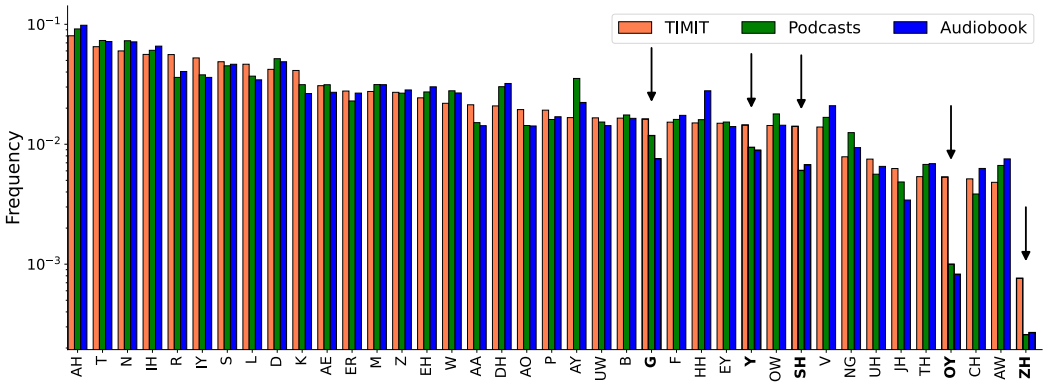

Figure 7: **Phoneme frequency distribution per corpus**. Bar plots show normalised phoneme frequencies for each corpus, sorted by decreasing frequency in TIMIT. Each phoneme category on the x-axis contains three adjacent bars corresponding to TIMIT (orange), Podcasts (green), and Audiobook (blue). The y-axis is displayed on a logarithmic scale to enhance visibility of low-frequency phonemes. Selected low-frequency phonemes (G, Y, SH, OY, ZH) are highlighted with downward-pointing arrows and thicker bar outlines to facilitate comparison across corpora.

leads to a substantially broader coverage of acoustic–phonetic patterns. From the perspective of MEG-based neural decoding, this increased variability provides a stronger test of whether models can maintain performance across both shifts in phoneme statistics and differences in their realisation, and whether increased diversity and scale facilitate the learning of representations that remain effective across heterogeneous linguistic inputs.

Finally, at the semantic level, the three corpora differ in the richness and diversity of their lexical content, providing a higher-level dimension of variability. The audiobook corpus is relatively constrained, as it centres on Sherlock Holmes narratives, leading to redundancy in settings, characters, and thematic structure. TIMIT, by contrast, spans a broad range of topics but consists of short, largely unrelated sentences, limiting coherence across samples. The podcast corpus provides the greatest semantic diversity, comprising multiple extended narratives on distinct topics. These differences are illustrated in Fig. 9, where a pretrained Word2Vec model (Mikolov et al., 2013) was used to extract word embeddings for representative keywords. The embeddings were normalised and projected into two dimensions using cosine distance. In semantic space, keywords show a compact cluster for the audiobook corpus and more dispersed distributions for the TIMIT and, especially, the podcasts corpus. More generally, the corpora occupy partially distinct regions of the semantic embedding space, reflecting differences in lexical content. Taken together, they provide broader semantic coverage than any individual corpus alone. This diversity allows us to examine how decoding models perform across variations in meaning and discourse context, and whether exposure to semantically richer and more varied inputs supports more robust performance. In combination with the other sources of variability, this also enables us to investigate how scale and diversity jointly influence the stability of learned neural representations in naturalistic speech perception.

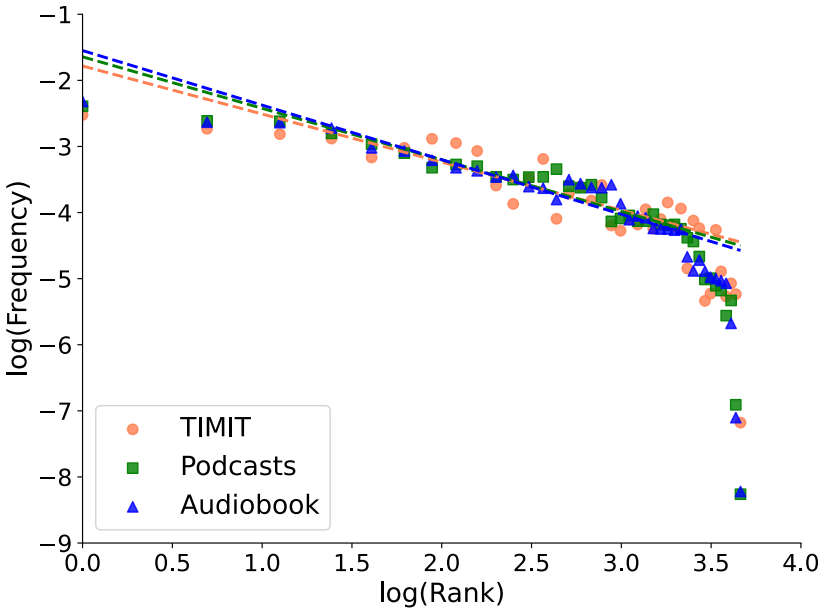

Figure 8: **Phoneme rank-frequency distribution per corpus**. Each point represents a phoneme, with its position determined by its rank (x-axis) and normalised frequency (y-axis), both shown in logarithmic scale. Phonemes are ranked in decreasing order of frequency based on TIMIT. Dashed lines indicate linear fits computed over ranks. Different marker shapes (circles, squares, triangles) and different colors (orange, green, blue) distinguish the three corpora.

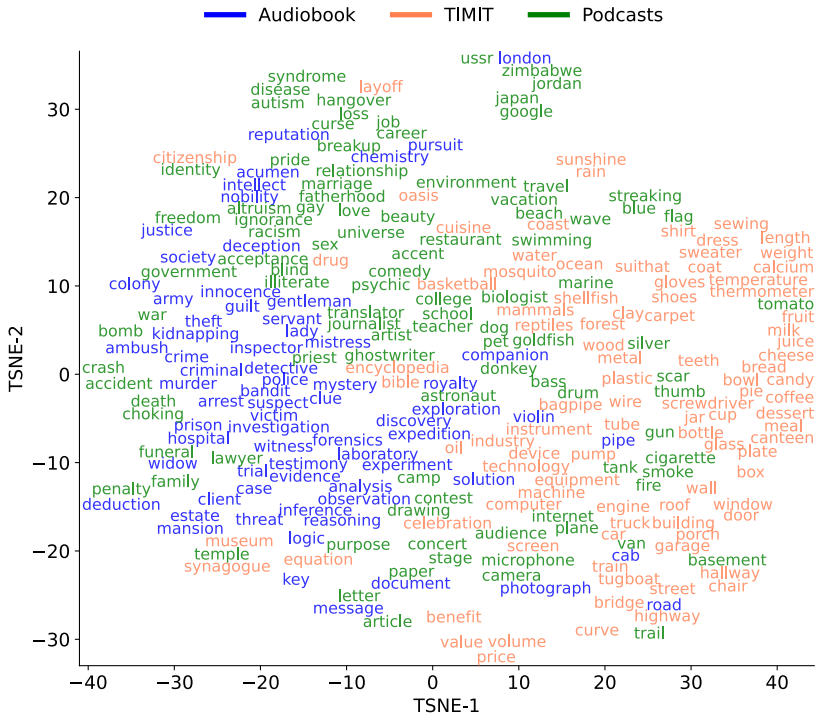

Figure 9: **Keyword embeddings per corpus**. Two-dimensional t-SNE projection of keyword embeddings from the Audiobook (blue), TIMIT (orange), and Podcasts (green) corpus . Each point corresponds to a keyword, and proximity between points reflects similarity in the embedding space, capturing lexical-semantic relationships between words.

D Additional MEG Details

D.1 Neural Variability

A challenge in decoding speech perception from MEG lies in the variability of the neural signal itself. Unlike controlled experimental paradigms, naturalistic listening introduces fluctuations that arise not only from the stimulus, but also from the listener. In the present dataset, this variability is intrinsic to the design: recordings span multiple participants and multiple sessions per participant, each associated with differences in attention, engagement, fatigue, and individual neurophysiology. Participants vary in language background (native vs. non-native), in their level of sustained attention to the narrative, and in cognitive states such as drowsiness or mind wandering. Even within the same individual, these factors can fluctuate across recording sessions, especially when sessions are conducted consecutively, leading to changes in alertness and engagement. In addition, inter-individual differences in brain anatomy, head shape, and age-related factors influence the spatial configuration and magnitude of the measured MEG signals. Together, these sources of variability shape the observed neural responses and constitute a fundamental challenge for any model aiming to extract robust neural representations of speech.

To characterise this variability, we examined three complementary dimensions of the MEG signal — amplitude, phase, and power — contrasting speech and non-speech segments. These measures capture distinct aspects of auditory cortical processing. The root mean square (RMS) of the evoked response reflects the strength of stimulus-locked activity, which is typically enhanced over bilateral temporal sensors during speech perception. Inter-trial coherence (ITC) quantifies the consistency of phase alignment across trials, and is known to increase at low frequencies when neural activity entrains to the temporal structure of speech. Finally, power spectral density (PSD) provides a frequency-resolved measure of oscillatory activity, with speech perception commonly associated with increased power in the delta–theta frequency band (~2–7 Hz), corresponding to prosodic and syllabic rhythms (Peelle and Davis, 2012).

We first assessed the consistency of these neural signatures within a single participant across multiple recording sessions (Figure 10). The evoked responses reveal clear commonalities: speech segments elicit stronger amplitude over bilateral temporal regions, increased phase alignment at low frequencies, and elevated low-frequency power for speech segments as compared to non-speech segments. These effects are robust at the group level and consistent with established findings in auditory neuroscience. However, when examining individual sessions in detail (Figure 12), substantial variability emerges. The magnitude of amplitude differences, the degree of phase locking, and the extent of low-frequency power enhancement vary noticeably across sessions. Some sessions exhibit pronounced and clean separations between speech and non-speech, whereas others show attenuated or noisier patterns. This within-subject variability likely reflects fluctuations in attention, fatigue, and engagement with the narrative.

A similar pattern is observed across participants (Figure 11, 13), where shared neural signatures coexist with inter-individual variability. At the group level, the expected response patterns associated with speech perception are preserved: increased evoked amplitude over bilateral auditory cortices, stronger low-frequency phase coherence, and enhanced delta–theta power for speech relative to non-speech segments. Yet, the expression of these effects differs considerably between individuals. Some participants show strong and spatially focal responses, while others exhibit weaker or more diffuse patterns. These differences can arise from a combination of anatomical variability, differences in signal-to-noise ratio, age-related factors, and variability in cognitive engagement. Behavioural measures, such as responses to comprehension questions, further suggest that not all participants maintain the same level of attention throughout the recordings, which likely contributes to the observed heterogeneity in neural responses.

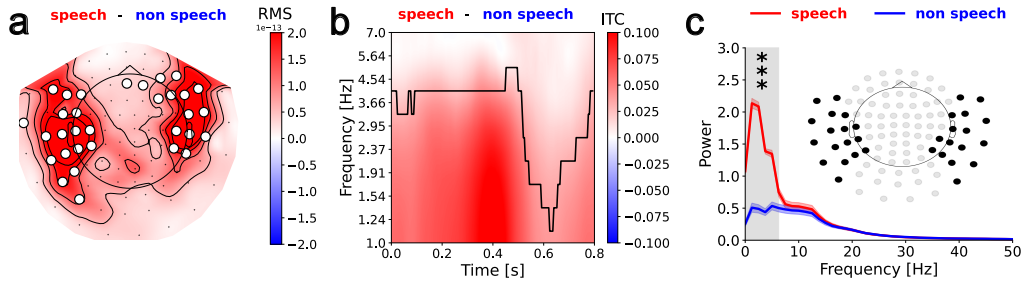

Figure 10: **MEG evoked response within subject**. Differences between speech and non-speech segments are shown as a proxy for auditory and linguistic processing. (**a**) Topographic map of the root mean square (RMS) difference between speech and non-speech evoked responses, averaged across sessions, showing stronger amplitude for speech over bilateral temporal sensors. Statistically significant sensors ($p<0.001$) are marked with white dots. (**b**) Time–frequency representation of the difference in inter-trial coherence (ITC) between speech and non-speech segments, averaged across sessions. Speech elicits greater phase consistency in low-frequency bands (~2–7 Hz; delta–theta range). Statistically significant clusters ($p<0.001$) are outlined with a black contour. (**c**) Power spectral density (PSD) for speech (red) and non-speech (blue) segments, averaged across sessions, indicating increased low-frequency power during speech. Statistically significant frequency ranges ($p<0.001$) are highlighted with a semi-transparent grey region. For each panel, data are shown for a single representative participant (sub-0), averaged across recording sessions, where each session corresponds to one chapter (1-14) of *A Study in Scarlet* from Sherlock Holmes audiobook.

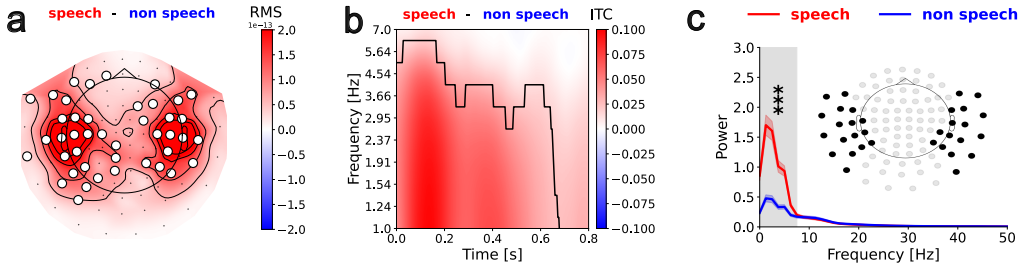

Figure 11: **MEG evoked response between subjects**. Data are shown for all participants (sub-0 to sub-32), averaged across subjects, for chapters 11–12 of *Study in Scarlet* from Sherlock Holmes audiobook. Differences between speech and non-speech segments are shown as a proxy for auditory and linguistic processing. (**a**) Topographic map of the root mean square (RMS) difference between speech and non-speech evoked responses, averaged across subjects, showing stronger amplitude for speech over bilateral temporal sensors. Statistically significant sensors ($p<0.001$) are marked with white dots. (**b**) Time–frequency representation of the difference in inter-trial coherence (ITC) between speech and non-speech segments, averaged across subjects. Speech elicits greater phase consistency in low-frequency bands (∼2–7 Hz; delta–theta range). Statistically significant clusters ($p<0.001$) are outlined with a black contour. (**c**) Power spectral density (PSD) for speech (red) and non-speech (blue) segments, averaged across subjects, indicating increased low-frequency power during speech. Statistically significant frequency ranges ($p<0.001$) are highlighted with a semi-transparent grey region.

Taken together, these analyses show that neural variability operates at multiple levels — both within and across subjects — and directly impacts the structure of the MEG signal used for decoding. Importantly, despite this variability, (Figures 10 and 11) reveal consistent and robust neural signatures across sessions and participants, including increased evoked amplitude over temporal regions, stronger low-frequency phase coherence, and enhanced delta–theta power during speech. These shared patterns indicate that a stable and meaningful signal is present across conditions, providing a common substrate that neural decoding models can exploit. At the same time, variability in the strength, spatial distribution, and reliability of these effects reflects fluctuations in attention, physiology, and recording conditions. As such, this variability is not merely noise to be eliminated, but a defining characteristic of the naturalistic speech perception task. It provides a critical testbed for evaluating the robustness

of deep learning models, which must learn to extract stable and behaviourally relevant features while remaining invariant to these sources of variation. The dataset therefore enables two complementary forms of generalisation: within-subject generalisation across sessions and linguistic contexts, and between-subject generalisation across individuals. By explicitly incorporating both shared structure and variability, it offers a realistic and challenging benchmark for developing models that can scale to the complexity of naturalistic speech perception.

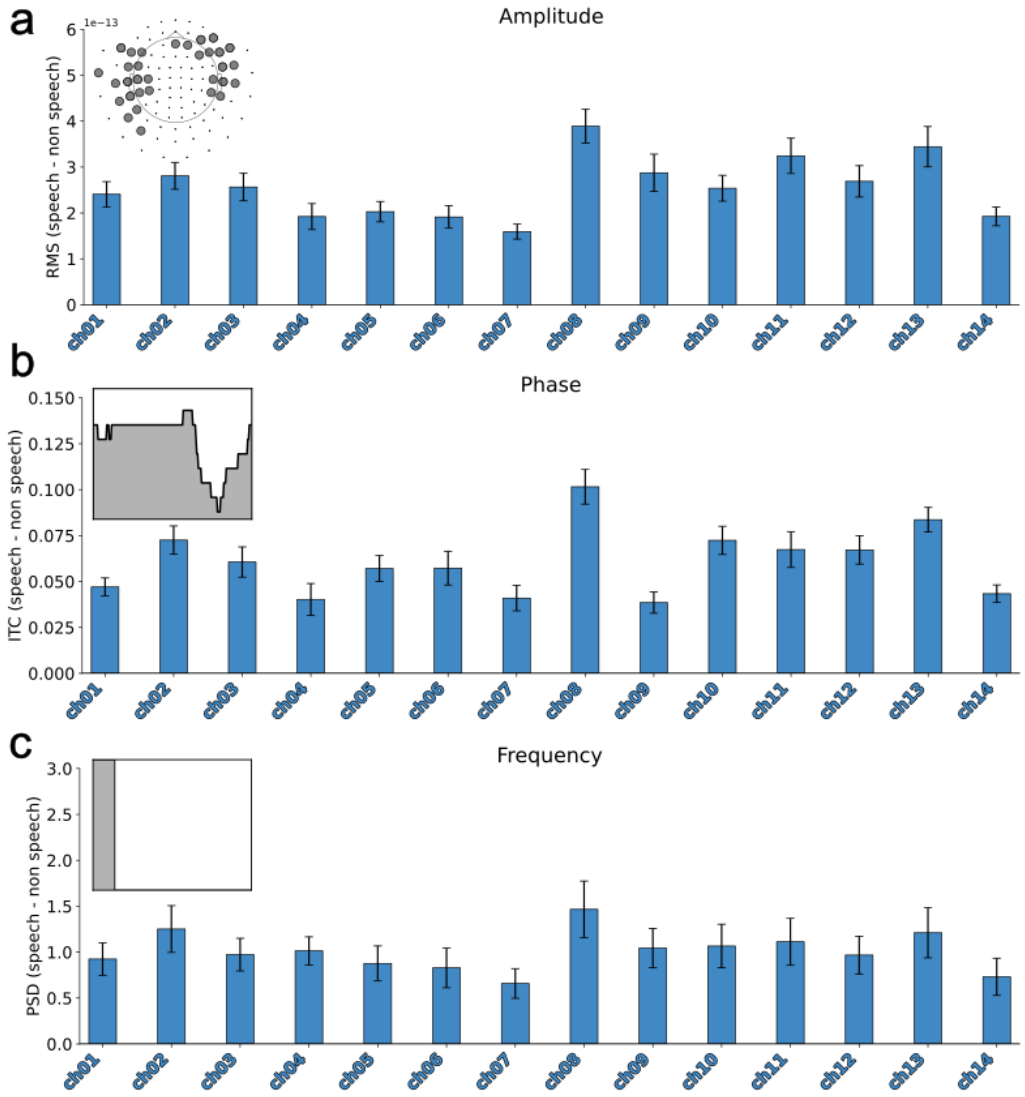

Figure 12: **Variability across recording sessions**. Spatial, temporal, and spectral components showing strong discrimination between speech and non-speech segments are illustrated across recording sessions to assess variability in amplitude, phase, and power. (**a**) Amplitude is quantified as the root mean square (RMS) of the evoked response. (**b**) Phase consistency is measured using inter-trial coherence (ITC). (**c**) Oscillatory power is measured using power spectral density (PSD). For each metric, the mean is computed within functionally relevant masks derived from statistical testing on aggregate data, shown as grey-shaded insets in each panel: a sensor mask for amplitude, a time–frequency mask for phase, and a frequency mask for power. Error bars indicate variability within the corresponding mask (across sensors, time–frequency points, or frequencies, depending on the panel). Data are shown for a single representative participant (sub-0), with each bar corresponding to one recording session, i.e., one chapter (1–14) of *A Study in Scarlet* from Sherlock Holmes audiobook.

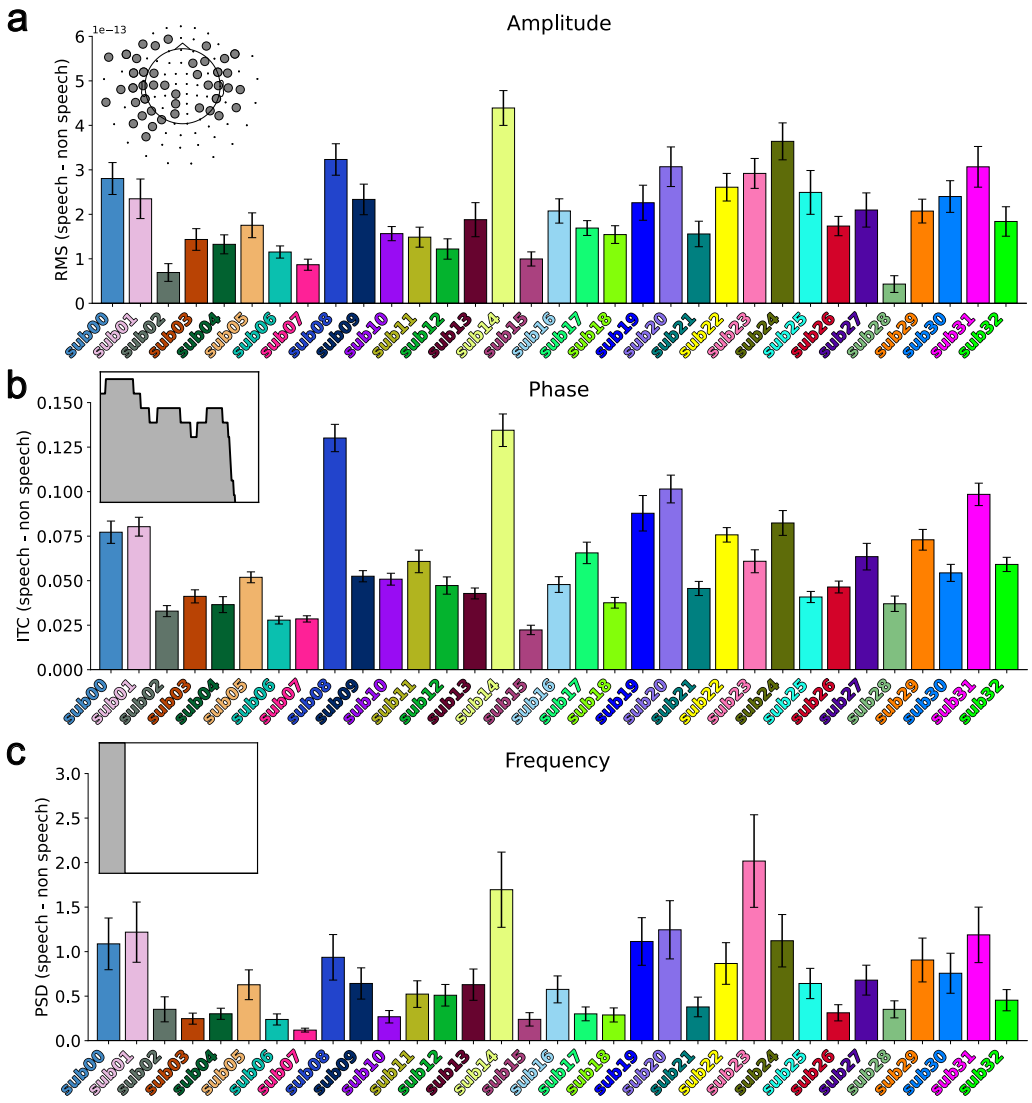

Figure 13: **Variability across subjects**. Spatial, temporal, and spectral components showing strong discrimination between speech and non-speech segments are illustrated across participants to assess variability in amplitude, phase, and power. (**a**) Amplitude is quantified as the root mean square (RMS) of the evoked response. (**b**) Phase consistency is measured using inter-trial coherence (ITC). (**c**) Oscillatory power is measured using power spectral density (PSD). For each metric, the mean is computed within functionally relevant masks derived from statistical testing on aggregate data, shown as grey-shaded insets in each panel: a sensor mask for amplitude, a time–frequency mask for phase, and a frequency mask for power. Error bars indicate variability within the corresponding mask (across sensors, time–frequency points, or frequencies, depending on the panel). Data are shown across participants (sub-0 to sub-32), with each bar corresponding to one subject. Colours differentiate participants.

E Additional Decoding Experiments

E.1 Generalisation of Supervised and Pre-trained Models to Deep and Broad Data

Building on the deep within-subject data of LibriBrain, LibriBrain100 introduces a new *breadth* axis of scaling through 32 new subjects with relatively little within-subject data compared to Subject 0. These two regimes appear to favour different kinds of models, with supervised models performing best on the deep data portion (Subject 0), while fine-tuning a self-supervised model pre-trained

over many subjects performs best for generalisation over Subjects 1-32 (Figure 14). This reflects the findings in Jayalath and Parker Jones (2026), who observe that with sufficient (deep) data, the statistical priors of pre-trained methods are superseded by rich in-domain data, whereas on shallow multi-subject data, the pre-trained multi-subject priors assist generalisation.

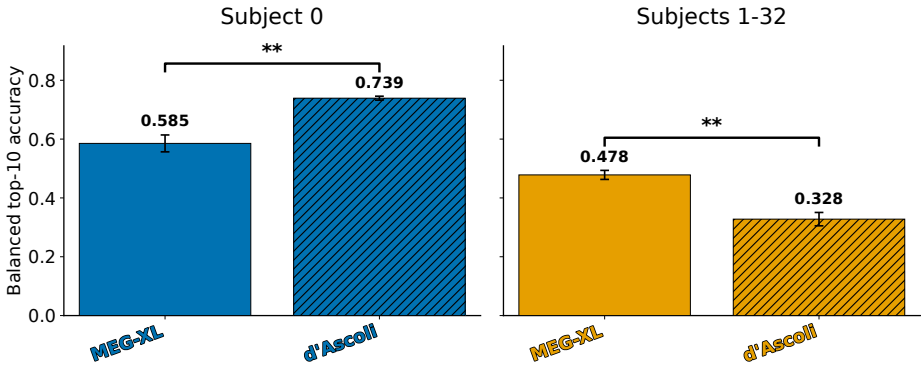

Figure 14: **d'Ascoli (Supervised) vs MEG-XL (Self-supervised).** We compare the supervised word decoding model proposed by d'Ascoli et al. (2025) to fine-tuning the pre-trained self-supervised MEG-XL model proposed by Jayalath and Parker Jones (2026). On deep Subject 0 data, the supervised model performs better, while on the shallow data of subjects 1-32, MEG-XL generalises better owing to its data-efficiency. Random chance is 0.2. ** indicates $p < .01$ under a Mann-Whitney U-test.

E.2 Measuring Information Transfer with LibriBrain100

The eventual hope in releasing LibriBrain100 is that it will accelerate progress towards non-invasive brain–computer interfaces that restore speech to paralysed patients. To this end, we benchmark the information throughput of models trained on LibriBrain100. We measure OVMI (Jayalath et al., 2026), an information-theoretic quantity for the mutual information between a user's intent and a decoding model, indexed to a particular communication distribution. We compare results on LibriBrain100 to that of a prior surgically implanted speech BCI used on a patient with anarthria (Moses et al., 2021). The early landmark result in Moses et al. (2021), with a simple 50-word vocabulary, paved the way for significant advances in following years (Willett et al., 2023; Card et al., 2024). By measuring performance against Moses et al. (2021), we aim to track how far the non-invasive decoding field is from reaching a similar breakthrough moment. The results in Figure 15 should be read with caution, however. LibriBrain100 is a heard speech dataset rather than an intended speech dataset, so the reported throughput does not represent that of a practical communication interface which would require evaluation against attempted or imagined speech decoding. Nevertheless, the results are indicative of significant progress being made towards strong performance on heard speech decoding through non-invasive methods.

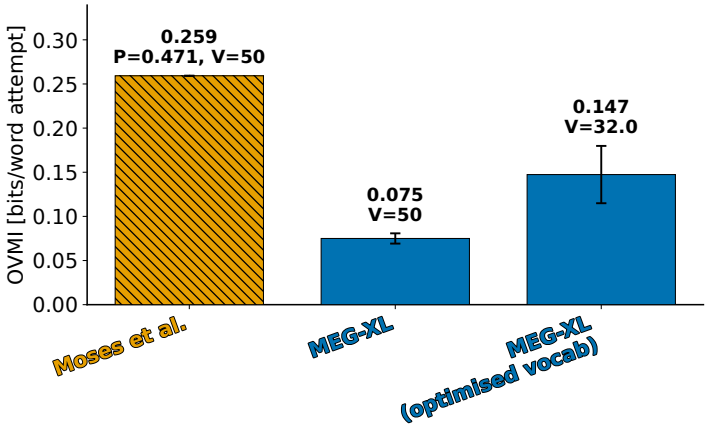

Figure 15: **Information transfer comparison.** We measure the OVMI achieved by Moses et al. (2021), calculating the quantity from their reported 47.1% decoder accuracy and 50-word vocabulary, against MEG-XL fine-tuned on LibriBrain with our target vocabulary and against an OVMI-optimised vocabulary. We use SUBTLEX-UK as the communication distribution p that parameterises OVMI. Higher OVMI scores are better. The results here are provided as references for future benchmarks.

F Additional Decoding Experiment Details

F.1 Computational Requirements

All experiments were performed on individual NVIDIA H100 GPUs on a system with 64 GiB of CPU memory. Fine-tuning MEG-XL with 100 hours of data took approximately 20 hours per run.

F.2 Target Words

The 50-word vocabulary was designed to span both high-frequency content and function words. It includes pronouns (e.g., “she”, “him”, “i”, “we”), conjunctions and prepositions (e.g., “and”, “but”, “on”, “at”), determiners and quantifiers (e.g., “the”, “a”, “any”), negation (e.g., “not”), auxiliary and modal verbs (e.g., “was”, “is”, “will”, “can”), common lexical verbs (e.g., “think”, “do”), and a small number of content words (e.g., “people”, “time”, “good”, and “new”) (Table 6).

Even with only 50 words, the vocabulary can support a range of short, practically useful utterances in the context of assistive communication. Examples include state reports (“i am good”, “i am not good”, “i think this is good”, “i can do that”, “i am out of it”), requests (“can i have that”, “can he be there”, “do not do that”), location or attention cues (“i will be there”, “is it time”), and social reference (“she was good to me”).

Table 6: **The 50-word vocabulary used in the word classification task**.

#	Word	#	Word	#	Word	#	Word	#	Word
1	is	11	he	21	but	31	at	41	do
2	the	12	that	22	will	32	out	42	can
3	a	13	have	23	so	33	our	43	time
4	to	14	this	24	all	34	am	44	think
5	it	15	they	25	my	35	it’s	45	good
6	i	16	of	26	for	36	had	46	always
7	not	17	there	27	she	37	him	47	new
8	was	18	and	28	were	38	an	48	people
9	we	19	are	29	any	39	very	49	as
10	be	20	in	30	really	40	has	50	on

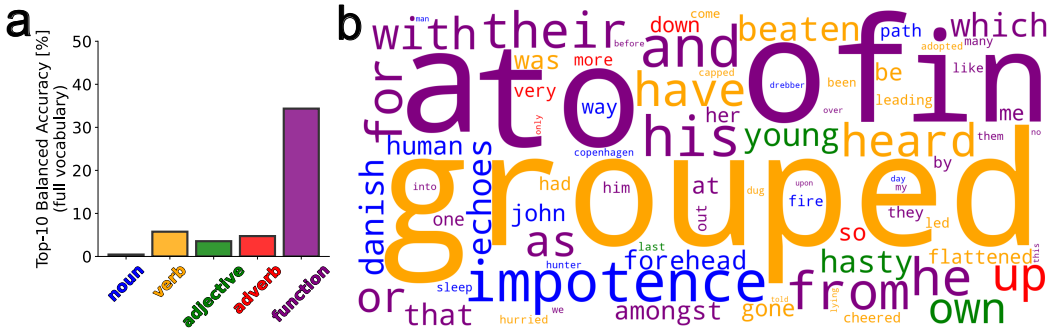

Figure 16: **Model predictions on full vocabulary**. (**a**) Top-10 accuracy across all word occurrences in Chapters 11 and 12 of A Study in Scarlet and across all 32 subjects, grouped by Part-of-Speech (POS) category (purple function, red adverb, green adjective, orange verb, blue noun). Predictions were evaluated against the full vocabulary, and accuracy is defined as the proportion of events for which the true word appears within the top-10 predicted words. (**b**) Word cloud of the 100 best-predicted words, ranked by top-10 balanced accuracy. Word size is proportional to top-10 balanced accuracy, and word colours indicate their corresponding Part-of-Speech categories.

G Ethical Considerations

The development and release of the LibriBrain100 dataset raise several important ethical considerations, which we have addressed throughout the research process:

Informed consent and participants privacy. All participants provided informed consent for data collection and explicitly approved the sharing of pseudoanonymised data for research purposes, in accordance with the University of Oxford's ethical oversight procedures. To safeguard privacy, all released data are subject to anonymisation procedures, including the removal or replacement of any information that could directly or indirectly identify individual participants. These steps are designed to minimise re-identification risk while preserving the utility of the dataset.

Open science and reproducibility. We intentionally rely on public-domain materials (e.g., LibriVox audiobooks and Project Gutenberg texts, TIMIT, MOCHA-TIMIT, The Moth Podcasts) and open-source tools, ensuring that the dataset can be freely redistributed without licensing restrictions. This commitment to openness extends to our analysis code, which is publicly available and documented to support reproducibility.

Dual-use considerations. While the primary motivation of this work is to support assistive communication technologies for clinical populations, brain decoding methods could be misused in ways that compromise privacy if applied without consent. Although current non-invasive approaches remain far from enabling such uses, we stress the importance of maintaining strong ethical standards, including informed consent and respect for participant autonomy. By releasing data and methods, we aim to support the development of shared ethical guidelines and responsible practices within the research community.

Long-term data stewardship. We are committed to ensuring the long-term availability and integrity of the dataset. The BIDS-formatted version will be hosted on established public platforms (e.g., OSF), and the dataset is also distributed via Hugging Face. Together with code hosted on GitHub, this provides redundancy and helps ensure sustained accessibility.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the properties of the dataset as described in Section 3, the experimental findings detailed in Section 4, and the limitations discussed in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: We do not present theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Aside from providing code, we describe all experimental settings, data splits, preprocessing steps, and evaluation protocols necessary to reproduce our results in Section 4, 4.1, 4.2, 4.3. We also ensure straightforward access to the dataset.
Our experiments build on previously published models (i.e., Jayalath and Parker Jones (2026) and d'Ascoli et al. (2025)). For these components, we refer to the original papers and publicly available implementations, which we explicitly reference. We clearly specify how these models are used and fine-tuned in our setting, including checkpoints, hyperparameters, and task definitions.
As our contribution focuses on introducing a new dataset and a comprehensive evaluation benchmark, we believe the combination of detailed experimental descriptions, dataset availability, and the use of standard, publicly available methods provides sufficient information to reproduce the main results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide execution-ready scripts to recreate the training runs that were used to produce the experimental results. The link to these is available in Section 3.3.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail information necessary to understand the main results in Section 4. We describe data splits in Table 1. We provide the full training, validation and test details in Section B.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars that represent standard error across multiple seeds and we mark statistically significant contrasts with asterisks in Figure 2, 3, 4 in Section 4. In Section D.1 we ran statistical tests (e.g., two relative sample t-test, cluster based permutation test) to assess the significance of the differences between speech and non-speech segments in amplitude, phase, and frequency (see Figure 10, 11)

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on computer resources need to reproduce the results in Appendix F.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We discuss ethical considerations at length in a dedicated Section G. The NeurIPS Code of Ethics was reviewed and adhered to.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss how we commit to prevent potential negative societal impacts concerning especially data privacy in Section G.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There are no immediate risks for misuse associated with the data and models we are publishing. However, we are intentionally releasing them under a CC BY-NC license to retain the ability to prevent potential unethical commercial use in the future.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: When we used existing software (e.g., Gentle, Praat) we appropriately credited the authors in the main text and explicitly referenced the package and its version.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide extensive documentation alongside our code and dataset on GitHub and Hugging Face including information on licensing, training, and limitations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: In Section B.3 and B.4 we give a detailed account of instructions given to participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: All participants provided informed consent for data collection and approved the sharing of pseudonymised data for research purposes, in accordance with the University of Oxford's ethical oversight.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.