TangentFuse: Low-Latency MEG Speech Activity Detection via Riemannian Covariance–CNN Fusion

Aryan Mangla[1] , **Priyanshu Vij**[1] and **Raj Mehta**[2]
[1]Manipal Institute of Technology, Manipal, India
[2]Cornell Tech, Cornell University, New York, USA
mangla.aryan.333@gmail.com, priyanshuvij.work@gmail.com, rm985@cornell.edu

Abstract

Speech activity recognition in MEG-based non-invasive BCI systems provides a reliable speech gate that can trigger downstream decoders only when speech-related neural activity is present. Such a gate can help users interact with assistive devices in continuous settings. While MEG provides excellent temporal resolution, many MEG speech activity classifiers do not fully exploit available spatial information. We describe a hybrid MEG speech/non-speech classifier that combines a geometry-aware covariance branch with temporal neural streams. We compute shrinkage covariance matrices (SPD) and map them to a Riemannian tangent space around a reference mean; a logistic regression classifier operates on these features. A limited sensor array defines the region-of-interest component, while the final system adds a residual temporal fusion layer over aligned probability streams. The final decision rule, including fusion weights, thresholds, and sequence-level post-processing, is selected on validation only and then applied unchanged to frozen test. On a large within-subject MEG corpus, the final system achieved validation macro-F1 of 0.8972 and frozen-test macro-F1 of 0.8914. This provides a compute-efficient research prototype for MEG-based speech gating.

1 Introduction

The speech gate is an important and critical element of many non-invasive BCI (Brain Computer Interface) pipelines. In most cases, a speech gate is used to determine whether speech-related neural activity is present to warrant further processing. If the speech gate determines that there is insufficient signal, it can prevent unnecessary processing or output. The use of the speech gate can help assistive devices provide continuous and predictable results in a variety of real-world settings. Most MEG-based systems require the speech gate to make decisions quickly and accurately, including under class imbalance (speech vs non-speech), while remaining computationally light enough to fit low-latency constraints.

Why MEG. There are several reasons that MEG is attractive for this application and why developing a good speech gate is difficult. MEG provides millisecond temporal resolution and is sensitive to cortical population activity, which makes it ideal for detecting the rapid transitions between speech and non-speech states. Additionally, however, detecting the presence of speech related activity in very small time windows from a large number of sensors presents significant challenges. Sensor array recordings are often contaminated with artifacts and can vary significantly over time; the data distribution can be imbalanced; and "hard" examples cluster around speech onset/offset boundaries. Using standard preprocessing techniques (such as SSS/denoising, filtering, and standardized channel handling) is necessary but does not by itself resolve the modeling challenge [Gramfort and others, 2013; Gramfort and others, 2014]. A good speech gate needs to integrate robust feature representations with training methods that are resistant to imbalance and non-stationarity.

Many approaches to developing a speech gate fail to fully utilize cross-sensor correlation and covariance geometric properties. Speech processing is often closely associated with coordinated activity in auditory and temporal regions, including onset and sustained responses and slow-band entrainment to speech dynamics [Giraud and Poeppel, 2012; Hamilton *et al.*, 2018; Ding *et al.*, 2014]. Since these types of phenomena are inherently multi-variate (they manifest as co-activated patterns of sensors), a natural way to represent them is through the computation of the sensor-sensor covariance matrix for a given time window. While covariance matrices are SPD objects and the set of SPD matrices is not a flat Euclidean space; distances, averages and linearizations need to take into account the intrinsic geometry of SPD space to avoid deforming the relationships between variables that are important for classification [Barachant *et al.*, 2012; Barachant *et al.*, 2013; Congedo *et al.*, 2013; Congedo *et al.*, 2017]. Therefore, a natural approach would be to develop geometry-aware decoding pipelines that treat SPD covariances as first-class features.

Temporal deep models (e.g., CNN-based decoders) are widely used for non-invasive speech and language decoding from MEG/EEG signals [Dash *et al.*, 2019], including strong recent systems for non-invasive speech perception [Défossez *et al.*, 2023]. As a result, there is a rich literature of Riemannian approaches to covariance-based decod-

ing in EEG/BCI that emphasize the importance of robustness and data efficiency through manifold respecting metrics and tangent space mappings, with recent surveys and extensions of these approaches [Tibermacine *et al.*, 2024; Andreev *et al.*, 2025]. There are also a growing number of deep learning approaches that exploit the SPD structure directly [Huang and Van Gool, 2017]. Rather than contributing to this theoretical area, the contribution of this paper lies in the development of a practical, low latency, fusion architecture that integrates a compact temporal CNN with an explicit geometry aware SPD covariance stream that are motivated by the unique information captured by each.

Contributions. We introduce TangentFuse, a MEG speech activity detector that (i) integrates a compact temporal CNN with a geometry aware SPD covariance branch based on shrinkage covariance estimation and Riemannian tangent space features; (ii) utilizes simple, defensible choices for stabilizing the detector, including an auditory/temporal ROI component, silence-aware sampling, and label-smoothed class-weighted BCE [Gramfort and others, 2013; Gramfort and others, 2014]; and (iii) selects the operating point of the final residual temporal-fusion system on validation macro-F1, including lightweight calibrators and sequence-level post-processing, before reporting frozen-test performance. The remaining portion of this paper provides additional detail regarding the complete architecture, places it in context with prior work and evaluates the detector on a large, within subject MEG corpus for speech activity detection [Ozdogan *et al.*, 2025].

2 Related Work

Speech Detection As Front End For Non-Invasive Decoding. Non-invasive speech decoding has made rapid progress through the use of large datasets and better sequence models. Most recently, researchers have found that decoding pipelines often benefit from a "speech active vs. inactive" decision that serves as a filter (gate) for downstream, more complex decoders. Especially when working with continuous signals and short time windows, identifying speech boundaries (onset/offset) is often among the most difficult parts of the task, with errors concentrated around onset/offset transitions [Défossez *et al.*, 2023].

Benchmark Context And Within Subject Scale. The LibriBrain dataset and the PNPL 2025 Competition formalize speech activity detection and other related tasks at a level that support much more systematic comparisons and protocol standardization within subject MEG settings [Ozdogan *et al.*, 2025; Landau and others, 2025]. These benchmarks highlight several aspects of practical importance for the problem of making a gating decision: short time windows, large class imbalances between speech and non-speech and the need to accurately detect transitions. Therefore, our evaluation is structured similarly to this benchmark framework and views speech activity detection as an operational front end versus an end to end speech content decoder.

The Repeated Use Of Covariance In Speech Decoding From MEG/EEG Data. A common theme in research on

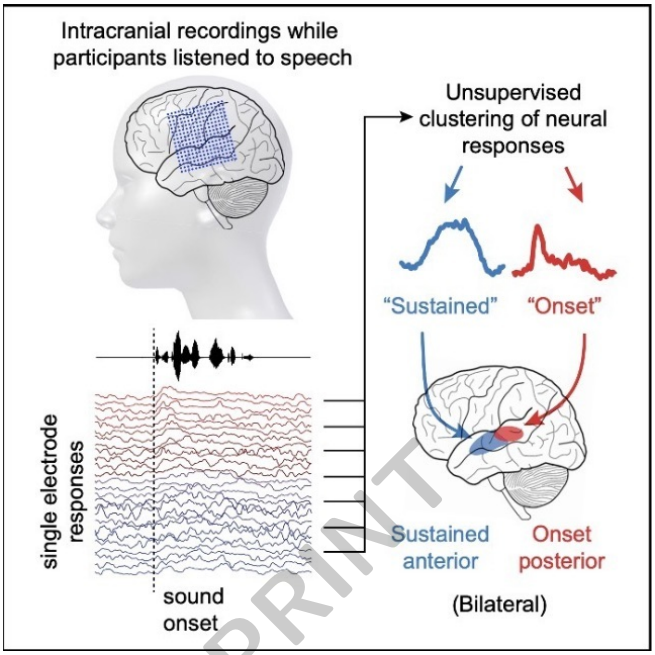

Figure 1: Spatial organization of onset and sustained responses in the human superior temporal gyrus during speech perception. Onset-selective regions are posterior, while sustained responses are more anterior. Adapted from Hamilton et al. (2018).

speech neuroscience is that speech processing is distributed and coordinated. Onset and sustained responses occur in multiple areas of the superior temporal region (see Figure 1) and slow timescale entrainment to continuous speech implies multi-sensor processing rather than independent processing per sensor [Giraud and Poeppel, 2012; Hamilton *et al.*, 2018; Ding *et al.*, 2014]. Therefore, covariance (or other second order statistics) is a natural representation of cross-sensor coordination within short time frames. Thus, pipelines that model sensor/sensor relationships are motivated as a complementary approach to temporal feature extraction from raw waveforms.

Covariance matrices are symmetric positive semidefinite (PSD) in general. When in short-window/high-dimensional regimes the sample covariance can be rank-deficient. Shrinkage is commonly used to obtain a well-conditioned positive-definite (SPD) estimate suitable for Riemannian distances and tangent-space features. Riemannian approaches utilize the geometric properties of the SPD manifold to define meaningful distances, averages, and linearizations for classification [Barachant *et al.*, 2012; Barachant *et al.*, 2013; Congedo *et al.*, 2013; Congedo *et al.*, 2017]. Traditional approaches include Minimum Distance To Mean (MDM) and tangent space mappings that map SPD matrices onto a Euclidean space where simple linear models are efficient and computationally tractable. Recent surveys and extensions have documented that such geometry aware pipelines are often robust in noisy environments, with small sample sizes and distribution shifts common to EEG/MEG recordings [Tibermacine *et al.*, 2024; Andreev *et al.*, 2025]. Our geometry aware pipeline follows the same formula (shrinkage covariance $\rightarrow$ tangent space $\rightarrow$ linear classifier) and claims no novel Riemannian methodol-

ogy; the novel contribution is in how this pipeline is coupled with a simple temporal model under latency constraints.

Estimation of covariance and shrinkage in high-dimensional settings with limited sample size per window. Pragmatic covariance based decoding relies heavily on estimating well conditioned covariance matrices from limited samples per window. Naively using sample covariance can lead to unstable estimates in these regimes. Shrinkage estimators, including Ledoit-Wolf type approaches and related MMSE shrinkage formulations, are widely used to stabilize covariance estimation in high dimensional settings [Ledoit and Wolf, 2004; Chen *et al.*, 2010] and motivated our selection to use shrinkage covariance estimation prior to tangent space mapping to keep the geometry aware component lightweight and improve numerical stability and downstream classification reliability.

There is an increase in literature on deep learning components that are made to work with SPD structure, including Riemannian networks for learning SPD matrices and end to end deep Riemannian networks for EEG decoding [Huang and Van Gool, 2017; Wilson *et al.*, 2022]. Similar concepts are being explored in computer vision through global covariance pooling and matrix normalization layers that allow for second order representations within deep architectures [Li *et al.*, 2018]. These approaches are important because they show that SPD structure can be incorporated into deep architectures, but they also add complexity and computational requirements to training. In contrast, TangentFuse is a practical fusion recipe: a simple temporal CNN captures discriminative waveform motifs, while a separate geometry aware covariance stream captures cross sensor coordination and the two are fused at the last stage of processing for low latency inference without needing to perform end to end SPD back propagation [Huang and Van Gool, 2017; Wilson *et al.*, 2022].

When moving across sessions or subjects, methods of tangent space alignment provide a principled means of reducing domain mismatch by aligning covariance geometry across domains [Bleuzé *et al.*, 2022]. Although the majority of our evaluations are for within subject detection, this literature provides support for the broader claim that tangent space representations can be engineered for robustness and portability, and suggests future extensions of our covariance stream to cross session or cross subject gating.

When working with imbalanced detection problems, it may be difficult to compare models by using a single or fixed threshold, and real-world applications will have to select the optimal operating point for the particular application of interest, depending on how they want to balance the trade-offs of their target metric and costs. Therefore, we developed a validation-based approach to select the optimal operating point: first, we calibrated the streams for temperature scaling and second, we selected the late fusion weights and the final decision thresholds for the best performance on the macro F1 score for the validation set. The macro F1 score treats both classes as equally important and is appropriate when both false positives and false negatives matter, which is common when onset/offset windows are hard. This aligns with current classifier practice that separates the probability calibration from the selection of the decision threshold in modern classifiers, but provides a very easy and reproducible way of doing this.

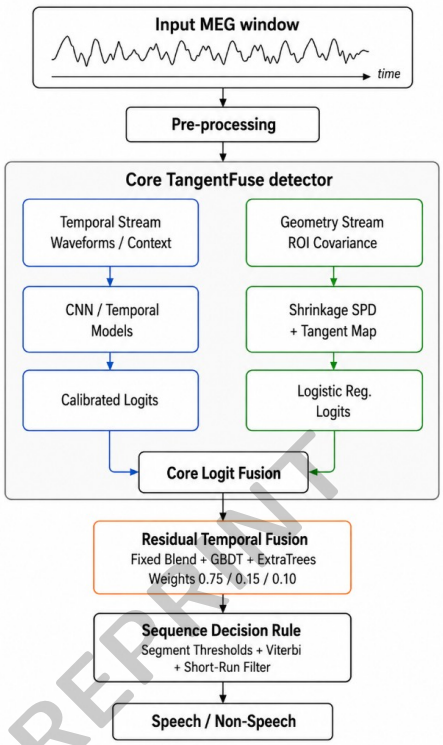

Figure 2: TangentFuse architecture overview. The core detector processes MEG windows with two complementary streams: (1) a temporal neural stream operating on MEG waveforms, and (2) a geometry-aware stream that computes shrinkage covariance, maps to tangent space, and applies logistic regression. The final reported system adds a validation-selected residual temporal-fusion layer over aligned probability streams, lightweight GBDT/ExtraTrees calibrators, and sequence-level post-processing.

3 Method

TangentFuse is a fusion framework for MEG speech activity detection (see Figure 2). Its core detector contains two complementary streams. First, a compact temporal CNN operates on windows of sensor time series. In parallel, a geometry-aware stream calculates a shrinkage covariance per window, maps it onto a Riemannian tangent space and classifies it with logistic regression. The core streams are fused by logit-level fusion with per-branch temperature scaling. The final reported system keeps this validation-driven fusion principle and adds a residual temporal-fusion layer over aligned probability streams to improve boundary and silence decisions.

3.1 Notations and Problem Setting

Let the continuous MEG recording of a subject be sampled at frequency f_s (here $f_s = 250$ Hz). The time series is divided into fixed-length windows of length W (here $W \approx 1$ s), resulting in $T = \lfloor W f_s \rfloor$ samples per window.

Each sample of the window index n, can be represented as the following MEG data matrix

$$\mathbf{X}^{(n)} \in \mathbb{R}^{C\times T}. \tag{1}$$

Where C denotes the number of selected MEG sensors (channels). Each row represents the time series of a single sensor during a window. The binary label is

$$y^{(n)} \in \{0, 1\} \tag{2}$$

Here $y^{(n)} = 1$ represents speech activity and $y^{(n)} = 0$ represents no speech activity.

In the task of learning a detector for speech activity, the goal is to determine a probability $\hat{p}^{(n)} \in [0, 1]$ of speech activity for each window and then produce a decision $\hat{y}^{(n)} \in \{0, 1\}$ through thresholding of this probability:

$$\hat{y}^{(n)}(t) = \mathbb{I}\big[\hat{p}^{(n)} \geq t\big]. \tag{3}$$

Since both types of errors have an impact on a speech gate (false positive errors trigger downstream processing; false negative errors suppress valid speech), we choose the operating threshold t based on the macro F1 score for the validation set, which is a common method of evaluating performance in this type of problem setting [Ozdogan *et al.*, 2025; Landau and others, 2025].

We denote the training, validation, and test sets by $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{va}}, \mathcal{D}_{\text{te}}$, where each of these three datasets consists of labeled windows $\{(\mathbf{X}^{(n)}, y^{(n)})\}$.

SPD and matrix-function notation. Let $\mathbb{S}^C_{++}$ be the set of $C \times C$ symmetric positive-definite (SPD) matrices. For any $A \in \mathbb{S}^C_{++}$, let its eigendecomposition be

$$A = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_C)\,U^\top, \quad \lambda_i > 0, \tag{4}$$

where U is orthonormal and $\{\lambda_i\}_{i=1}^C$ are the eigenvalues of A.

We define the matrix logarithm and matrix exponential (for SPD matrices) by acting on eigenvalues:

$$\log(A) = U\,\mathrm{diag}(\log\lambda_1, \ldots, \log\lambda_C)\,U^\top, \tag{5}$$

$$\exp(A) = U\,\mathrm{diag}(\exp\lambda_1, \ldots, \exp\lambda_C)\,U^\top. \tag{6}$$

Similarly,

$$A^{1/2} = U\,\mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_C^{1/2})\,U^\top, \tag{7}$$

$$A^{-1/2} = U\,\mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_C^{-1/2})\,U^\top. \tag{8}$$

These definitions are used explicitly in the tangent-space mapping in Section 3.4, avoiding ambiguity about the λ_i terms.

3.2 Preprocessing, Segmentation, and Sensor Subset

Standard MEG handling. It is assumed that the windows were extracted from the MEG data following the standard MEG pre-processing provided by most commonly used toolchains (e.g., MNE) which include standard channel orderings, and standard filter/denoising steps when possible [Gramfort and others, 2013; Gramfort and others, 2014]. Our method does not depend upon any special pre-processing beyond assuring that windows $\mathbf{X}^{(n)}$ are time-aligned, similarly scaled, and have no missing channels.

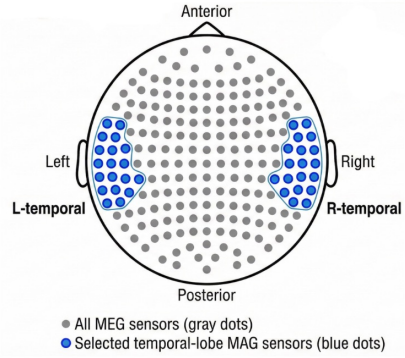

Figure 3: ROI sensor selection generated from the LibriBrain channel metadata and the fixed legacy ROI mask. Gray dots show all 306 MEG channels; blue dots show the 23 selected auditory/temporal ROI channels used by the ROI/covariance component.

Segmentation. A continuous record is segmented into W-second windows. Each window produces $\mathbf{X}^{(n)} \in \mathbb{R}^{C\times T}$ and its corresponding label $y^{(n)}$. No additional labeling rules are introduced; instead, the window-level labels provided by the dataset protocol are used [Ozdogan *et al.*, 2025; Landau and others, 2025].

Temporal-lobe sensor subset (ROI). Speech processing has been shown to be associated with the auditory/temporal regions [Giraud and Poeppel, 2012; Hamilton *et al.*, 2018; Ding *et al.*, 2014]. In order to focus the geometry-aware component on a relevant region of interest and decrease covariance dimensionality, the ROI analysis uses a predetermined set of sensors (see Figure 3). Specifically, the covariance branch uses $C = 23$ selected auditory/temporal MEG channels from the fixed ROI mask. The selected set contains both magnetometer and planar-gradiometer channels. The final residual temporal-fusion layer can additionally combine aligned probability streams from broader temporal models.

Silence-aware downsampling (reducing negative redundancy). The class distribution can be skewed, and long non-speech regions may contain a great deal of redundancy (i.e., long stretches of silence). To avoid over-emphasizing easy negative examples, we subsample negatives while keeping all positives. Let N_1 and N_0 be the count of positive and negative windows in the training split. Then, we retain each negative window with probability:

$$p_{\text{keep}} = \min\left(1,\ r\frac{N_1}{N_0}\right), \tag{9}$$

where $r > 0$ determines the desired ratio of negatives to positives after downsampling (for example, $r = 1$ would provide a roughly balanced ratio). This simple form of downsampling removes redundant silence windows, while still retaining enough negative examples near boundaries.

3.3 Temporal Stream

The temporal stream represents a compact 1-D convolutional classifier that maps a window $\mathbf{X}^{(n)} \in \mathbb{R}^{C\times T}$ to a single scalar

logit. Let $f_\theta(\cdot)$ represent the CNN, where θ are the CNN's parameters. The CNN's logit is

$$z_{\text{cnn}}^{(n)} = f_\theta\Big(\mathbf{X}^{(n)}\Big) \in \mathbb{R}, \qquad p_{\text{cnn}}^{(n)} = \sigma\Big(z_{\text{cnn}}^{(n)}\Big), \tag{10}$$

where $\sigma(u) = (1 + e^{-u})^{-1}$ is the logistic sigmoid. We use label smoothing and class weighting: let $\tilde{y}^{(n)} = (1-\epsilon)y^{(n)} + \epsilon/2$ with smoothing parameter $\epsilon > 0$, and let $\beta > 0$ be the positive-class weight selected from the training class ratio. The training loss is

$$\begin{aligned}\mathcal{L}_{\text{cnn}}^{(n)} = (1 - \tilde{y}^{(n)})\,\text{softplus}\Big(z_{\text{cnn}}^{(n)}\Big) \\ + \beta\,\tilde{y}^{(n)}\,\text{softplus}\Big(-z_{\text{cnn}}^{(n)}\Big).\end{aligned} \tag{11}$$

We optimize θ by minimizing the mean loss over the (downsampled) training set:

$$\min_\theta \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{(X^{(n)},y^{(n)})\in\mathcal{D}_{\text{tr}}} \mathcal{L}_{\text{cnn}}^{(n)}. \tag{12}$$

This combination (ROI sensors + silence-aware downsampling + smoothed/weighted BCE) constitutes the stabilizing configuration used for the temporal stream.

3.4 Geometry-Aware Stream

The geometry-aware stream represents each window by an SPD covariance matrix and performs classification in a Riemannian tangent space, following established covariance-decoding pipelines [Barachant *et al.*, 2012; Barachant *et al.*, 2013; Congedo *et al.*, 2013; Congedo *et al.*, 2017] with shrinkage covariance estimation for numerical stability [Ledoit and Wolf, 2004; Chen *et al.*, 2010].

Shrinkage Covariance Estimation
Given a window $\mathbf{X}^{(n)} \in \mathbb{R}^{C\times T}$, we remove the per-channel mean:

$$\boldsymbol{\mu}^{(n)} = \frac{1}{T}\,\mathbf{X}^{(n)}\mathbf{1}, \qquad \tilde{\mathbf{X}}^{(n)} = \mathbf{X}^{(n)} - \boldsymbol{\mu}^{(n)}\mathbf{1}^\top, \tag{13}$$

where $\mathbf{1} \in \mathbb{R}^T$ is the all-ones vector. We compute the sample covariance:

$$\mathbf{S}^{(n)} = \frac{1}{T}\,\tilde{\mathbf{X}}^{(n)}\tilde{\mathbf{X}}^{(n)\top}. \tag{14}$$

We adopt the $1/T$ normalization to match common shrinkage-covariance toolkits; the constant scaling does not affect the affine-invariant geometry used downstream. Following standard shrinkage constructions [Ledoit and Wolf, 2004; Chen *et al.*, 2010], we use the scaled-identity target:

$$\mathbf{F}^{(n)} = \frac{\text{tr}(\mathbf{S}^{(n)})}{C}\,\mathbf{I}, \tag{15}$$

and form a shrinkage covariance using oracle approximating shrinkage (OAS):

$$\boldsymbol{\Sigma}^{(n)} = (1 - \rho^{(n)})\,\mathbf{S}^{(n)} + \rho^{(n)}\,\mathbf{F}^{(n)}. \tag{16}$$

The shrinkage coefficient $\rho^{(n)}$ is computed using OAS, yielding a well-conditioned SPD matrix suitable for downstream geometry-aware operations.

Reference Mean
Let $\{\boldsymbol{\Sigma}^{(n)}\}$ denote training covariances. We define the reference SPD point as the Riemannian (Fréchet) mean under the affine-invariant metric [Barachant *et al.*, 2012; Barachant *et al.*, 2013; Congedo *et al.*, 2013; Congedo *et al.*, 2017]:

$$\bar{\boldsymbol{\Sigma}} = \arg\min_{\mathbf{G}\in\mathcal{S}_{++}^C} \sum_n \delta_{\text{R}}^2(\mathbf{G}, \boldsymbol{\Sigma}^{(n)}), \tag{17}$$

where $\delta_{\text{R}}(\cdot,\cdot)$ is the affine-invariant Riemannian distance. In practice, this mean is computed once from $\mathcal{D}_{\text{tr}}$ and held fixed for all subsequent mapping of validation and test covariances.

Tangent-Space Mapping
We map each covariance to the tangent space at $\bar{\boldsymbol{\Sigma}}$ using the whitened-log representation:

$$\mathbf{M}^{(n)} = \log\Big(\bar{\boldsymbol{\Sigma}}^{-1/2}\,\boldsymbol{\Sigma}^{(n)}\,\bar{\boldsymbol{\Sigma}}^{-1/2}\Big) \in \mathcal{S}_C. \tag{18}$$

We use the principal matrix logarithm; for SPD inputs this yields a symmetric matrix. We then vectorize the symmetric matrix with $\sqrt{2}$ scaling on off-diagonal entries:

$$\mathbf{v}^{(n)} = \text{vec}_{\text{sym}}\big(\mathbf{M}^{(n)}\big) \in \mathbb{R}^{C(C+1)/2}. \tag{19}$$

For $C = 23$, this yields $d = 276$ features.

Logistic Regression Classifier
A linear classifier produces a logit and probability:

$$z_{\text{ts}}^{(n)} = \mathbf{w}^\top\mathbf{v}^{(n)} + b, \qquad p_{\text{ts}}^{(n)} = \sigma\big(z_{\text{ts}}^{(n)}\big), \tag{20}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. We fit $(\mathbf{w}, b)$ on $\mathcal{D}_{\text{tr}}$ using logistic regression with ℓ_2 regularization [Pedregosa and others, 2011], handling imbalance via class weighting.

This completes the geometry-aware stream: shrinkage covariance → tangent-space mapping → logistic regression, following established Riemannian decoding pipelines.

3.5 Late Fusion

The two streams capture complementary information: the CNN learns discriminative temporal motifs from raw waveforms, while the tangent-space stream captures cross-sensor coordination encoded in SPD covariances. We fuse them at the logit level with calibration.

Per-Branch Temperature Scaling
Raw logits from different models are often on different scales; direct fusion can be dominated by the higher-magnitude logit stream. We therefore apply per-branch temperature scaling, tuned exclusively on $\mathcal{D}_{\text{va}}$. Let $\tau_{\text{cnn}} > 0$ and $\tau_{\text{ts}} > 0$ be temperatures. We define calibrated logits

$$\tilde{z}_{\text{cnn}}^{(n)} = \frac{z_{\text{cnn}}^{(n)}}{\tau_{\text{cnn}}}, \qquad \tilde{z}_{\text{ts}}^{(n)} = \frac{z_{\text{ts}}^{(n)}}{\tau_{\text{ts}}}. \tag{21}$$

Each temperature is selected on $\mathcal{D}_{\text{va}}$ by minimizing the negative log-likelihood (equivalently BCE) of that stream's predictions:

$$\begin{aligned}\tau_{\text{cnn}}^* = \arg\min_{\tau>0} \sum_{(X^{(n)},y^{(n)})\in\mathcal{D}_{\text{va}}} \\ \text{BCEWithLogits}\left(\frac{z_{\text{cnn}}^{(n)}}{\tau}, y^{(n)}\right),\end{aligned} \tag{22}$$

and similarly for τ_{ts}^*.

Logit-Level Fusion and Threshold Selection
The fused logit is a convex combination:

$$z_{\text{fuse}}^{(n)}(\alpha) = \alpha\, \tilde{z}_{\text{cnn}}^{(n)} + (1-\alpha)\, \tilde{z}_{\text{ts}}^{(n)}, \qquad \alpha \in [0,1]. \tag{23}$$

The fused probability is

$$p_{\text{fuse}}^{(n)}(\alpha) = \sigma\left(z_{\text{fuse}}^{(n)}(\alpha)\right). \tag{24}$$

We select α by grid search on $\mathcal{D}_{\text{va}}$. For each candidate α, we also select a decision threshold $t \in [0,1]$ to maximize macro-F1 on validation.

$$(\alpha^*, t^*) = \arg\max_{\alpha\in\mathcal{A},\, t\in\mathcal{T}} \text{MacroF1}\big(t;\, p_{\text{fuse}}(\alpha)\big), \tag{25}$$

where $\mathcal{A}$ and $\mathcal{T}$ are discrete grids (e.g., $\mathcal{A} = \{0, 0.05, \ldots, 1\}$, $\mathcal{T} = \{0, 0.01, \ldots, 1\}$). At test time, we report results using the fixed (α^*, t^*) found on validation.

Summary of the training/tuning protocol. (i) Train the temporal stream on $\mathcal{D}_{\text{tr}}$ using the stabilized objective in Section 3.3; (ii) compute shrinkage covariances and tangent features on $\mathcal{D}_{\text{tr}}$ and fit logistic regression; (iii) tune stream calibration, fusion weights, and thresholds on $\mathcal{D}_{\text{va}}$; and (iv) evaluate on $\mathcal{D}_{\text{te}}$ using the fixed validation-selected settings.

3.6 Residual Temporal Fusion

The final reported system uses a residual temporal-fusion layer to correct errors left by the core window-level detectors. This layer combines time-aligned speech-probability streams with two lightweight train-fitted calibrators. Let z_{fix}, z_{gbdt}, and z_{et} denote the logits from the fixed temporal blend, a histogram-gradient-boosting calibrator, and an ExtraTrees calibrator. Here z_{fix} is the primary validation-selected temporal probability stream before adding the two tree-based residual calibrators; the tree calibrators are fitted on training-derived features and their weights are selected only on validation. Validation selected the final logit blend

$$z_{\text{final}} = 0.75 z_{\text{fix}} + 0.15 z_{\text{gbdt}} + 0.10 z_{\text{et}}. \tag{26}$$

The binary decision sequence is then obtained using validation-selected per-segment thresholds, a Viterbi transition penalty, and short-run filtering. No fusion weight, calibrator setting, threshold, or post-processing parameter is selected on the frozen test set.

4 Experiments

4.1 Dataset, Preprocessing, and Splits

To assess the speech activity detection task, we utilize the LibriBrain within-subject MEG corpus [Ozdogan *et al.*, 2025; Landau and others, 2025], which contains recordings from one participant (see Figure 4). The corpus includes 306 MEG channels (102 magnetometers, 204 planar gradiometers). As with all MNE-Python workflows [Gramfort and others, 2013; Gramfort and others, 2014], the recordings are first downsampled to 250 Hz and then segmented into fixed-length windows of about 1 second (250 samples). We do not employ a separate boundary class for this task; instead, windows were labeled as either silent or non-silent based upon the provided annotation protocol.

The data was divided into distinct training, validation, and frozen-test splits. In the local serialized inventory used for the final experiments, 91 runs were available; the paper-style split used 77 training runs, 7 validation runs, and 7 frozen-test runs. All model choices, fusion weights, thresholds, calibrator settings, and post-processing parameters were selected on validation only. The frozen-test split was evaluated only after the final configuration was fixed.

Metric	Validation	Frozen test
Macro-F1	0.8972	0.8914
Speech F1	0.9669	0.9616
Silence F1	0.8275	0.8212
Balanced accuracy	0.8978	0.8901

Table 1: Validation and frozen-test performance of the final validation-selected TangentFuse residual temporal-fusion system. Macro-F1 is the mean of speech F1 and silence F1.

4.2 Implementation Details

The core system has two parallel components. The temporal stream uses a compact CNN to process MEG windows. The CNN used a binary cross-entropy loss function modified to include label smoothing and class weighting calculated from the training class ratio. The geometry-aware stream processes each ROI window as a per-window SPD covariance matrix $\mathbf{\Sigma}^{(n)}$, utilizing an analytic shrinkage estimator (e.g., Ledoit–Wolf or OAS) following per-channel mean removal. After processing, these matrices are projected onto a Riemannian tangent space at a reference mean $\bar{\mathbf{\Sigma}}$ calculated from the training covariances, and the projected matrices are processed using logistic regression [pyRiemann Developers, 2025; Pedregosa and others, 2011].

Independent temperature scalars (τ_{cnn} and τ_{ts}) can be learned for each component during calibration of the outputs of each branch on the validation split. The final reported system further combines a fixed temporal blend with histogram-gradient-boosting and ExtraTrees calibrators using validation-selected logit weights of 0.75, 0.15, and 0.10. The final binary sequence uses validation-selected per-segment thresholds, Viterbi transition regularization, and short-run filtering.

4.3 Final Evaluation

The performance of our approach is evaluated using true macro-F1, defined as the mean of speech F1 and silence F1 (see Table 1). The final validation-selected TangentFuse residual temporal-fusion system achieved validation macro-F1 of 0.8972 and frozen-test macro-F1 of 0.8914. On frozen test, speech F1 was 0.9616, silence F1 was 0.8212, and balanced accuracy was 0.8901, indicating that most remaining errors are concentrated in silence and boundary regions rather than sustained speech regions. We report frozen-test and per-class metrics in addition to validation macro-F1 so that the operating point can be inspected beyond a single validation score.

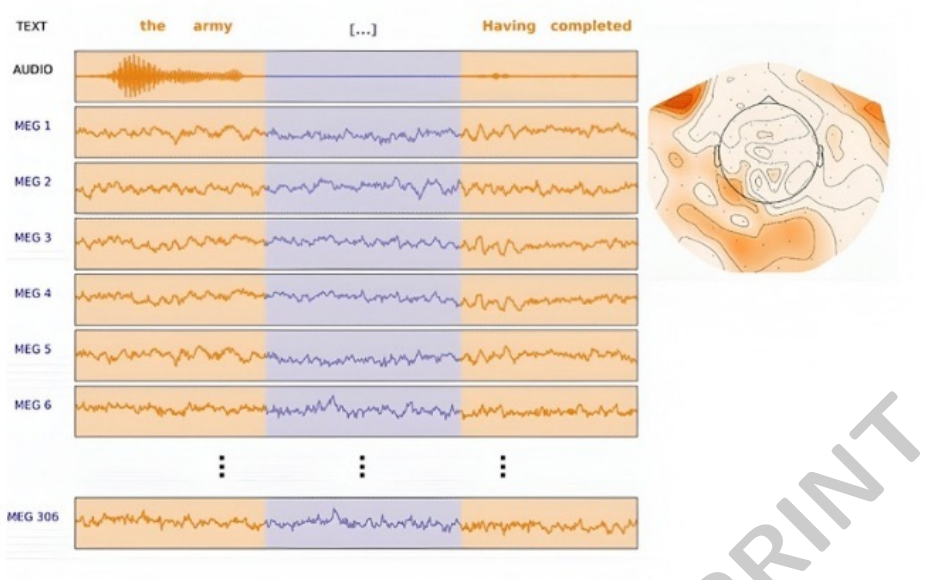

Figure 4: Illustration of MEG speech activity detection. Left: continuous recording showing text transcription, audio waveform, and selected MEG channels. Orange regions indicate speech activity; blue regions indicate silence. The task is to classify each 1-second window as speech or silence. Right: topographic map showing spatial distribution of MEG activity.

5 Discussion

The final results show that validation-selected fusion and temporal decision regularization are important for this task. Sustained speech regions are detected reliably, as shown by the frozen-test speech F1 of 0.9616. The lower silence F1 of 0.8212 suggests that the main bottleneck is not speech detection itself, but distinguishing silence and handling onset/offset boundaries.

The geometry-aware stream remains useful as a principled way to represent cross-sensor coordination: each window is summarized as a shrinkage SPD covariance estimate, mapped to a Riemannian tangent space about a reference mean, and then classified using a linear logistic regression model. The strongest empirical result, however, required an additional residual temporal-fusion layer. This layer combines aligned temporal probability streams with lightweight calibrators and sequence-level post-processing, which helps correct short false islands and boundary inconsistencies left by isolated window decisions.

For deployment-oriented use, the validation protocol is part of the model: all fusion weights, thresholds, calibrator settings, and sequence-level post-processing parameters were selected on validation only, and frozen test was evaluated using the fixed validation-selected configuration. The chosen operating point controls the trade-off between false triggers and missed speech, and should be reproduced with the same ROI, temporal context, and post-processing constraints.

6 Conclusion

In this paper we introduced TangentFuse, an MEG-based speech activity detection system using a combination of spatially-aware covariance features, temporal neural streams, validation-selected calibration, and residual temporal fusion. The final validation-selected configuration achieved validation macro-F1 of 0.8972 and frozen-test macro-F1 of 0.8914. These results suggest that temporal modeling, geometry-aware covariance features, and sequence-level calibration can be combined to create a viable within-subject MEG-based speech-gated research prototype. The low-latency framing refers to the short-window, lightweight design; hardware-specific latency measurements and cross-subject validation remain future work.

Ethical Statement

This is a research prototype and should never be considered or used as a clinically approved system or assistive technology until there have been further evaluations, appropriate validation, and the completion of all necessary regulatory requirements. Because MEG recordings are a type of sensitive biosignal, the use of these signals for any application must include informed consent from all users, and the data produced must be strictly controlled through appropriate access restrictions and proper handling of derived activity logs. The model's decision threshold and fusion parameters provide a mechanism to adjust the operational point of the model, but do not represent an "objective" or "absolute" measure of the ground truth; therefore, all deployments must document the parameters used to determine the operational point and assess the impact of that point (i.e., false alarms vs. missed speech) in the deployment environment. Possible misuse scenarios include using speech activity detection for surveillance purposes or to infer internal user state(s) beyond the stated assistive function; therefore, misuse must be explicitly prohibited,

and all deployments must be restricted to the minimum capabilities required to perform the stated BCI gating functions.

Acknowledgments

Current addresses: Aryan Mangla, NVIDIA; Priyanshu Vij, BlackRock. This work was performed before Aryan Mangla and Priyanshu Vij joined their current employers and was not performed as part of their employment at NVIDIA or BlackRock. The views and conclusions expressed are those of the authors and do not necessarily represent the views of NVIDIA, BlackRock, or their affiliates.

References

[Andreev *et al.*, 2025] Anton Andreev, Gregoire Cattan, and Marco Congedo. The riemannian means field classifier for EEG-based BCI data. *Sensors*, 25(7):2305, April 2025.

[Barachant *et al.*, 2012] Alexandre Barachant, Stephane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Trans. Biomed. Eng.*, 59(4):920–928, 2012.

[Barachant *et al.*, 2013] Alexandre Barachant, Stephane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, 2013.

[Bleuzé *et al.*, 2022] Alexandre Bleuzé, Jérémie Mattout, and Marco Congedo. Tangent space alignment: Transfer learning for brain–computer interface. *Frontiers in Human Neuroscience*, 16:1049985, 2022.

[Chen *et al.*, 2010] Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero III. Shrinkage algorithms for MMSE covariance estimation. *IEEE Trans. Signal Process.*, 58(10):5016–5029, 2010.

[Congedo *et al.*, 2013] Marco Congedo, Alexandre Barachant, and Anton Andreev. A new generation of brain–computer interface based on Riemannian geometry. *arXiv preprint arXiv:1310.8115*, 2013.

[Congedo *et al.*, 2017] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain–computer interfaces: A primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

[Dash *et al.*, 2019] Debadatta Dash, Paul Ferrari, and Jun Wang. Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. In *Proc. IEEE EMBC*, pages 5531–5534, 2019.

[Défossez *et al.*, 2023] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5:1097–1107, 2023.

[Ding *et al.*, 2014] Nai Ding, Monita Chatterjee, and Jonathan Z. Simon. Cortical entrainment to continuous speech: Functional roles and relations to speech intelligibility. *Frontiers in Human Neuroscience*, 8:311, 2014.

[Giraud and Poeppel, 2012] Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517, 2012.

[Gramfort and others, 2013] Alexandre Gramfort et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013.

[Gramfort and others, 2014] Alexandre Gramfort et al. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, 2014.

[Hamilton *et al.*, 2018] Liberty S. Hamilton, Erik Edwards, and Edward F. Chang. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, 28(12):1860–1871.e4, 2018.

[Huang and Van Gool, 2017] Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. In *Proc. AAAI*, 2017.

[Landau and others, 2025] Gilad Landau et al. The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. *arXiv preprint arXiv:2506.10165*, 2025.

[Ledoit and Wolf, 2004] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[Li *et al.*, 2018] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proc. IEEE/CVF CVPR*, 2018.

[Ozdogan *et al.*, 2025] Murat Ozdogan, Gilad Landau, Greg Elvers, Dineth Jayalath, Pranav Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *arXiv preprint arXiv:2506.02098*, 2025.

[Pedregosa and others, 2011] Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[pyRiemann Developers, 2025] pyRiemann Developers. pyRiemann documentation (v0.7): "tangentspace" and "fgmdm" classes. https://pyriemann.readthedocs.io/en/0.7/, 2025. Accessed September 2025.

[Tibermacine *et al.*, 2024] Ibtissam Elhassani Tibermacine, S. Russo, A. Tibermacine, A. Rabehi, B. Nail, K. Kadri, and C. Napoli. Riemannian geometry-based EEG approaches: A literature review. *arXiv preprint arXiv:2407.20250*, 2024.

[Wilson *et al.*, 2022] Daniel Wilson, Robin T. Schirrmeister, Lukas A. W. Gemein, and Tonio Ball. Deep Riemannian networks for end-to-end EEG decoding. *arXiv preprint arXiv:2212.10426*, 2022.