

Whole-brain dynamics of articulatory, acoustic and semantic speech representations

Corresponding Author: Ms Maxime Verwoert

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Communications Biology.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Title: Whole-brain dynamics of articulatory, acoustic and semantic speech representations

Verwoert and colleagues carried out a very interesting investigation on the neural signatures of speech production in the case of single-word production. sEEG data was recorded from 16-60 year old individuals as they utter 100 single-words. The sEEG channels have a broad bilateral coverage. The results indicate more widespread and temporally dynamic contributions from the left compared to the right hemisphere. The relationship between speech and sEEG was assessed by considering articulatory, acoustic, and semantic features. In doing so, the authors could identify the selectivity of different neural sites to specific features, substantially building a spatial selectivity map. The results indicate a bilateral engagement in the speech production process, with some left dominance.

This is certainly interesting data. I found the analyses to be simpler than what the text would lead to believe. I think it is necessary to revisit the write-up quite a bit, adding clarification especially regarding the methodology, and maybe recalibrate some of the observations. But the data and results are potentially interesting and, in my opinion, are worth publishing.

Major comments

- I think the manuscript lacks clarity on the dependent measure. I was forced to carefully read the methods before getting into the results, as the introduction and result do not provide any broad intuitions on what the metrics actually are. And even when I studied the methods section, I found missing details that the authors must revisit (please see specific points in my minor comments). I found particularly unclear the explanation of the linear regression. I think this is an issue with the writing, but I will need to see the revised version to be fully convinced of that.
- Many methodological choices were not motivated in the text. For example, starting from the experimental paradigm: why single-words? Why linear regression? Why decoding rather than encoding? In fact, encoding models could consider all features simultaneously while, in my understanding, the current analyses focusses on one feature and one channel at a time. See my specific points below, where I highlight other parts that are unclear.
- Regarding the result in Fig. 2B. Isn't it obvious that articulatory and acoustic encodings have a large overlap? Maybe I missed that, but it sounds likely to me that the features themselves are highly correlated (while that would not be the case with semantics). Have the authors calculated how well one feature can be predicted from another with the same method? This would affect the overlap for the mapping in Figure 3 and the corresponding conclusions. But maybe the correlation is not as "bad" as I think?
- I find the title a bit deceiving. From what I gather, the main result is the spatial map that is obtained with the linear regression, not the study of temporal dynamics, which looks secondary to me. So, my view is that either the text does not sufficiently highlight the importance of the findings about temporal dynamics or that, more likely, the main finding is about the spatial mapping of these features. In that case, I suggest rewording the title to reflect the main finding.

Minor comments

- Abstract: "complementary". Complementary to what?
- "Intracranial speech production data": Unclear. This should state "intracranially recorded neural data recorded during speech production", or something like that.

- A small note. Is the term “While-brain” in the title appropriate? Maybe “distributed” would be more suitable (as you mention on line 52)? Even if the coverage is impressive, whole brain sounds excessive. Also, it might be better if the title reflected the finding (that you identified a pathway or a distributed processing etc), rather than on the methodology (broad coverage with iEEG).
- Line 19: The authors make it sound as if deep learning enabled the study of speech in naturalistic scenarios. That is not the case. In fact, deep learning is often used for feature extraction (as oppose as for probing the neural signal), and the key methodologies for probing the neural processing of continuous speech does not really require deep learning (for example, see work with non-invasive neural recordings from Edmund Lalor, Jonathan Simon, and Liberty Hamilton).
- Word2Vec is sufficient for a single-word paradigm. But have the authors considered using embeddings from more modern models? I found that the text lacks motivations of this kind.
- Line 20: The authors claim that there has been a shift from encoding to decoding models. That’s an unnecessary statement that I find unsubstantiated. The use of both encoding and decoding (and encoding-decoding) models have been rising in recent years. Decoding models are useful, but they are limited when it comes to interpretability. Of course, they are ideal when we talk about BCIs. Encoding models are widely used, especially when trying to account for multiple features at once (e.g., the multivariate Temporal Response Function – Crosse et al., *Front in Hum. Neurosci.*, 2016).
- Line 63: The text should clarify what “significant channels” means in this context. I understand the need to keep this brief, but the manuscript jumps into the results without even a brief explanation of what these metrics might be. That part has to be revised.
- Figure 2: does red reflect all channels sensitive to articulatory features, or just the ones that are uniquely sensitive to articulatory features? (indeed, the methods were named at the end of the introduction, but nothing is said about the dependent measures).
- Line 248: More details should be included on that filtering procedure.
- Lines 247 and 256: It’s unclear if an anti-aliasing filter was applied as part of these downsampling and decimation operations.
- Line 258: Add a few sentences to explain the procedure. Also, the code could use further high-level explanations in the header.
- Line 272: I find it confusing that you explain three times the same preprocessing procedure for the neural data. Is there a reason? Am I missing something?
- Line 268: I am not sure I understand. You are preprocessing and aligning the stimulus feature and the neural signal. Then, you perform a linear regression to quantify the relationship between the two. You say that you are “predicting” articulatory features for each channel. Does that mean that you are “decoding” articulatory features from a single channel? Does that include any time-latencies, or is this model only measuring instantaneous relationships? i.e., what do the authors mean exactly by “each timeframe”?
- Data sharing: I gave the data on OSF a quick look. Is the neural data in there? The documentation is insufficient. There should be at least a document explaining what all the files represent.

Reviewer #2

(Remarks to the Author)

This study uses the sEEG system to investigate where articulatory, acoustic, and semantic representations are processed in the brain. The articulatory, acoustic, and semantic representations were estimated based on Acoustic-to-Articulatory Inversion Recurrent Neural Network, compressed speech spectrogram, and semantic embeddings from Word2Vec respectively. The study aimed to identify significant channels that contributed to processing of one of these representations or a combination of these representations. The study concluded that both hemispheres contributed to processing of all representations, however, there was greater involvement of the left (compared to the right hemisphere). The semantic representations were processed by widely distributed channels, which were distinct from the channels responsible for processing articulatory and acoustic representations. The conclusions provide converging evidence on localization of articulatory, acoustic, and semantic representations.

I have four major issues with the study at least in the form in which it is currently presented.

- 1) In the abstract and throughout the introduction the readers are promised that this study will “limit potential con-founders” or help “acquire crucial complementary information for speech BCI’s”. I do not see how the current results bring us closer to be able to use BCI’s or how this study helps to “open the path for speech neuroprostheses”. Perhaps a more in-depth discussion of how the results of this study helps with this aim could be added to the discussion. Additionally, I wonder to what extend mapping of representations for individuals with unimpaired speech will help decode speech in individuals that experience speech deficits due to brain impairments (e.g. stroke, TBI)?
- 2) The way the methodology is currently described doesn’t allow for replication. Crucial information on how the representations were calculated is difficult to find or missing. For example, you mention “21 parameters represent trajectories of different articulatory muscles”. Are these parameters pre-specified by the pipeline, or were these selected for this experiment? It is mentioned on line 268 that “We trained a linear regression in a 10-fold cross-validation... The folds were averaged to obtain the final score per channel, timeframe and articulator”. Where exactly can I find these results, because I did not see results per articulator in the paper. Could you elaborate on the use of semantic embeddings in the linear regression? In general, the whole methodology should be made more detailed.
- 3) I am wondering about the appropriateness of the statistical tests used. The later sections of the results (anatomical contributions to speech representations, temporal dynamics of speech representations), have a “descriptive character” and I

am missing some statistics based on which the conclusions can be drawn. This is especially true for statements like: "the semantic representation appears to be stronger in the right hemisphere in these primary regions specifically". I am also wondering whether drawing the main conclusions based on t-tests is the most appropriate. Does other literature in this field use this approach?

4) I find that one of your important findings is not discussed in relation to other literature anywhere in the discussion (see p. 4: "The temporal dynamic results indicate that the majority of articulatory- and/or acoustic-selective channels are active well before speech onset... This could reflect the communication between different areas for speech preparation")

Minor Issues:

1) I worry that the readers might misinterpret the following sentence: "there appears to be a relatively stronger tuning for articulatory features in the left hemisphere versus a relatively stronger tuning for acoustics in the right hemisphere (Fig. 2C)". Some might read this as "there is stronger tuning for acoustics in the right hemisphere in general". Could you perhaps rephrase?

2) I am not in agreement with the interpretation of this null result "This revealed no difference between the left (38.47 ± 16.69) and the right (37.25 ± 17.57) hemisphere overall ($t(1645) = 1.44$, $p = 0.15$ ". Bayesian statistics would have to be used for the claim of "no difference".

3) The section titles "Anatomical contributions to speech representations" are too general, as this study only discusses distinct aspects of speech representations.

4) When grouping results into 3 time segments. Why did you opt for segments of different lengths?

5) The time windows for the semantic representations are not clear from the text (p. 3).

6) I would like figure S1 to be a lot clearer. Could you perhaps include one figure per participant, so the readers can get a better idea about electrode placements?

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed the issues I raised carefully. They provided very thorough explanations and justifications of their chosen analyses and writing style. The methods and results are a lot clearer in the current version. There are a couple of minor clarifications that could benefit the reader in the beginning of the result section:

1. On line 89: "For all three speech representations (articulatory, acoustic and semantic), there are significant channels present in both hemispheres (Fig. 2A)." It may be helpful to briefly reiterate how the significant channels were calculated to aid reader comprehension.

2. On line 113: "Only for one individual the articulatory-selective channels were spread farther than the semantic ones (Fig. S2)." Figure S2 shows the results separately per hemisphere (which is informative for the readers), however, it is not immediately clear that statement above refers to averages across both hemispheres.

The current study provides valuable contribution to studying neural dynamics of various stages in the speech production process. I look forward to seeing the manuscript published.

Reviewer #3

(Remarks to the Author)

This revision has addressed my concerns. Well done on an excellent study.

I have a couple of minor comments (the current version can be published as is, in my view).

- Lines 82-85: The authors use a valid method, and the explanation is sufficient. I find it somewhat unconvincing in that there is no explicit mention to why the same things could not be done with an encoding models (I can imagine the motivation, but it's not sufficiently explicit in the text). Nonetheless, I find the explanation sufficient for a generalist reader.

- Regarding the title (my final major comment from the previous review). We agree to disagree. I still think that the results of this paper are primarily on the spatial dimension. So, a title reflecting that would be more appropriate in my view. It's not a deal-breaker for me, so I'll leave this to the editor.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers

Reviewer #1 (Remarks to the Author):

Title: Whole-brain dynamics of articulatory, acoustic and semantic speech representations

Verwoert and colleagues carried out a very interesting investigation on the neural signatures of speech production in the case of single-word production. sEEG data was recorded from 16-60 year old individuals as they utter 100 single-words. The sEEG channels have a broad bilateral coverage. The results indicate more widespread and temporally dynamic contributions from the left compared to the right hemisphere. The relationship between speech and sEEG was assessed by considering articulatory, acoustic, and semantic features. In doing so, the authors could identify the selectivity of different neural sites to specific features, substantially building a spatial selectivity map. The results indicate a bilateral engagement in the speech production process, with some left dominance.

This is certainly interesting data. I found the analyses to be simpler than what the text would lead to believe. I think it is necessary to revisit the write-up quite a bit, adding clarification especially regarding the methodology, and maybe recalibrate some of the observations. But the data and results are potentially interesting and, in my opinion, are worth publishing.

We thank the reviewer for their time, suggestions and kind words about our manuscript. We agree with the reviewer that clarifications would improve the manuscript and will address each specific point below.

With regards to the simpler analyses than expected, while we could certainly apply more complex models (e.g., a recurrent encoder-decoder (Kohler et al., 2022, <https://doi.org/10.51628/001c.57524>)), we preferred to use simple analyses here for easy interpretation. The goal was to compare the representations, not to optimize the outcome. We have added this motivation to the latter part of the introduction to make the nature of the analyses clear so that reader would not be led to believe differently: [lines 82-85] "These relatively simple analyses enable a clear interpretation of the spatial and temporal dynamics of these speech representations, focusing on uncovering overlapping and distinct anatomical patterns involved in speech production rather than optimizing correlational strength or predictive power."

Major comments

- I think the manuscript lacks clarity on the dependent measure. I was forced to carefully read the methods before getting into the results, as the introduction and result do not provide any broad intuitions on what the metrics actually are. And even when I studied the methods section, I found missing details that the authors must revisit (please see specific points in my minor comments). I found particularly unclear the explanation of the linear regression. I think this is an issue with the writing, but I will need to see the revised version to be fully convinced of that.

We appreciate the careful reading of the manuscript. We agree that more information is necessary to make a good transition from the introduction to the results section without reading the methods. To accommodate this, we have added an explanation of the metrics in the last paragraph of the introduction:

[lines 78-82] "Next, to examine the neural distribution of these speech representations, we correlate the neural activity of each electrode channel with the extracted articulatory and acoustic features and use a linear regression encoding model to estimate neural

activity across trials from the extracted semantic embeddings. We then identify which channels and points in time encode semantic information and which correlate with articulatory or acoustic features significantly above chance level.”

- Many methodological choices were not motivated in the text. For example, starting from the experimental paradigm: why single-words? Why linear regression? Why decoding rather than encoding? In fact, encoding models could consider all features simultaneously while, in my understanding, the current analyses focusses on one feature and one channel at a time. See my specific points below, where I highlight other parts that are unclear.

We apologize if our methods are not motivated sufficiently. These choices are motivated by earlier research in the field and within our laboratory, as well as a desire to keep things simple to allow for highly interpretable results. The use of a single-word paradigm allows us to look at word embeddings without influence of the broader context. This broader context is known to influence neural activity substantially, as for example shown in recent work looking at neural firing rates during pronouns (Dijksterhuis et al., 2024, DOI: [10.1126/science.adr281](https://doi.org/10.1126/science.adr281)). We have added this motivation to the 'Task' section in the methodology:

[lines 265-269] “The choice of a single-word paradigm was motivated by the goal to generate highly interpretable results and to minimize complexity. Full sentences would add further semantic dependencies that are beyond the scope of this study. Additionally, single words provide an optimal framework for the articulatory-to-acoustic inversion model that is used to generate the articulatory features, by reducing the complexity of coarticulation effects and anticipatory gestures present in continuous speech.”

We focus on one feature and channel (and time window) at a time specifically to be able to look at the spatial (and temporal) differences. We have tried to clarify this consideration better in the introduction with this line:

[lines 82-85] “These relatively simple analyses enable a clear interpretation of the spatial and temporal dynamics of these speech representations, focusing on uncovering overlapping and distinct anatomical patterns involved in speech production rather than optimizing correlational strength or predictive power.”

While we reformulated the linear regression to simple correlations for the articulatory and acoustic representations (explained below), we did keep linear regression in the semantic pipeline. We chose this as it is a simple and fast model, a standard in the field, and allows for interpretable results. We believe this to be sufficient since the goal is not to optimize the outcome. Linear regression has also been used in the previous studies investigation semantics described in the introduction and also referred to in this section: “This procedure closely follows previous studies on semantic representations \cite{goldstein2022shared,herff2023semantic}.”. We have added another sentence to highlight the motivation:

[lines 365-367] “Linear regression was chosen for its interpretability and suitability for the relatively small number of word trials, allowing us to capture the linear relationship between the semantic features and neural activity.”

- Regarding the result in Fig. 2B. Isn't it obvious that articulatory and acoustic encodings have a large overlap? Maybe I missed that, but it sounds likely to me that the features themselves are highly correlated (while that would not be the case with semantics). Have the authors calculated how well one feature can be predicted from another with the same method? This would affect the overlap for the mapping in Figure 3 and the corresponding conclusions. But maybe the correlation is not as “bad” as I think?

The reviewer is correct in that articulatory and acoustic features are naturally correlated, as specific articulatory movements are required to produce particular sounds. However, these features capture different aspects of the speech production process. For example, producing the word "*mama*" involves continuous bilabial movements (lips) for the /m/ sound and limited tongue movement. Similarly, the word "*sigh*" requires minimal tongue movement, as it primarily involves airflow through the teeth (fricative /s/) and the jaw opening for the vowel /ai/, yet the acoustic properties are quite distinct. These examples highlight that articulatory features represent physical muscle movements, while acoustic features encode the resulting sound properties, even if they overlap to some extent. Another reason is the fact that in our work, the articulatory features are derived from the acoustic data (through the acoustic-to-articulatory inversion model). However, these features (acoustic and articulatory) still represent different processes and are often used (separately) in the speech BCI literature. Therefore, the potential neural overlap between them in terms of spatial and temporal dynamics (and thus potential implant locations for a BCI) was expected and of particular interest to us. We have attempted to highlight their correlation better in the text.

In the introduction:

[lines 45-47] "Another approach is **at the acoustic level, the sounds that are produced through movement of the articulatory muscles**. Spectrotemporal features **can be extracted from** recorded speech audio to obtain an acoustic representation of speech."

In the discussion:

[lines 178-179] "The articulatory and acoustic representations share a similar neural distribution, most prominently within and surrounding the Sylvian fissure. **A high degree of overlap between these features can be expected due to their natural correlation**, and has similarly been found in earlier studies..."

- I find the title a bit deceiving. From what I gather, the main result is the spatial map that is obtained with the linear regression, not the study of temporal dynamics, which looks secondary to me. So, my view is that either the text does not sufficiently highlight the importance of the findings about temporal dynamics or that, more likely, the main finding is about the spatial mapping of these features. In that case, I suggest rewording the title to reflect the main finding.

With 'whole-brain dynamics' we intended a double interpretation of both spatial (how the representations are spatially different/dynamic between one another) and temporal (how the representations change over time) dynamics. We think this is described well in the abstract and the introduction and would therefore like to keep the original title.

Minor comments

- Abstract: "complementary". Complementary to what?

We agree that this phrase was unfortunately worded. We meant complementary to the more common approach of using articulatory 'motor-based' features to decode speech (for a BCI). In combination with other comments about this sentence and the guideline to shorten the abstract, we have decided to remove this sentence from the abstract entirely.

- "Intracranial speech production data": Unclear. This should state "intracranially recorded neural data recorded during speech production", or something like that.

Indeed, we have corrected this sentence in the following way:

[lines 7-8] "Intracranial neural data was recorded from 15 participants, across 1647 electrode contacts, while they overtly spoke 100 unique words."

- A small note. Is the term "Whole-brain" in the title appropriate? Maybe "distributed" would be more suitable (as you mention on line 52)? Even if the coverage is impressive, whole brain sounds excessive. Also, it might be better if the title reflected the finding (that you identified a pathway or a distributed processing etc), rather than on the methodology (broad coverage with iEEG).

While we appreciate the concern and suggestion from the reviewer, we believe that "whole-brain" represents the extensive distributed coverage well in concise wording that would be appropriate for a short and clear title. Along with the previous argumentation regarding the use of "dynamics", we would like to keep the original title.

- Line 19: The authors make it sound as if deep learning enabled the study of speech in naturalistic scenarios. That is not the case. In fact, deep learning is often used for feature extraction (as oppose as for probing the neural signal), and the key methodologies for probing the neural processing of continuous speech does not really require deep learning (for example, see work with non-invasive neural recordings from Edmund Lalor, Jonathan Simon, and Liberty Hamilton).

We thank the reviewer for catching this, we did not write this sentence clearly. We indeed meant the use for feature extraction, in analyzing the acoustic rather than neural data. We have changed this sentence to reflect this more clearly:

[lines 33-36] "Advances in Natural Language Processing and Deep Neural Networks now enable data-driven approaches that can more efficiently extract complex features of acoustic speech, especially in naturalistic settings."

- Word2Vec is sufficient for a single-word paradigm. But have the authors considered using embeddings from more modern models? I found that the text lacks motivations of this kind.

We use the word2vec model because it provides an effective semantic representation while being well-suited for tasks constrained to single words, as it does not rely on a larger contextual horizon. Given our single-word paradigm, we believe that a more complex model would not offer additional benefits. We have clarified this motivation in the 'Semantic representation' part of the methods section:

[lines 360-362] "The word2vec model was chosen for its simplicity and suitability for single-word representations, avoiding the contextual complexity of more modern models that are unnecessary for single words."

- Line 20: The authors claim that there has been a shift from encoding to decoding models. That's an unnecessary statement that I find unsubstantiated. The use of both encoding and decoding (and encoding-decoding) models have been rising in recent years. Decoding models are useful, but they are limited when it comes to interpretability. Of course, they are ideal when we talk about BCIs. Encoding models are widely used, especially when trying to account for multiple features at once (e.g., the multivariate Temporal Response Function – Crosse et al., Front in Hum. Neurosci, 2016).

The reviewer again makes a valid point, we have approached this paragraph too much from a BCI perspective. We have attempted to rectify it by removing the unnecessary statement about encoding-decoding and instead make the transition to introduce BCI in the following way:

[lines 38-42] "These methods complement traditional approaches by providing deeper insights into the complex dynamics of spoken language and have also facilitated progress in the development of speech neuroprostheses, or, brain-computer interfaces (BCIs). Speech BCIs aim to restore communication in individuals who have lost their ability to speak by decoding neural signals and transforming it into a reconstruction or classification of speech features."

- Line 63: The text should clarify what "significant channels" means in this context. I understand the need to keep this brief, but the manuscript jumps into the results without even a brief explanation of what these metrics might be. That part has to be revised.

The reviewer is absolutely right, we hope that this is now clear with the added information in the introduction:

[lines 81-82] "We then identify which channels and points in time encode semantic information and which correlate with articulatory or acoustic features significantly above chance level."

- Figure 2: does red reflect all channels sensitive to articulatory features, or just the ones that are uniquely sensitive to articulatory features? (indeed, the methods were named at the end of the introduction, but nothing is said about the dependent measures).

The red color represents the channels uniquely sensitive (significant) to articulatory features in figure 2A, since channels sensitive to multiple representations have a unique color (yellow, magenta, cyan or light gray). However, in figures 2B-C it is the total number of channels correlating with articulatory features. We have clarified this in the description of figure 2A:

[lines 124-125] "Locations of channels significantly correlated with each representation uniquely (red - articulatory, green - acoustic, blue - semantic) or with overlapping representations (colors are disseminated in the Venn diagram). Smaller gray dots represent non-significant channels."

- Line 248: More details should be included on that filtering procedure.

We have added the following details:

[lines 309-311] "We first processed the audio by manually **silencing** background noises unrelated to speech **using Audacity. It was not necessary to apply sophisticated filtering to separate speech from noise, since noises did not temporally coincide with speech utterances.**"

- Lines 247 and 256: It's unclear if an anti-aliasing filter was applied as part of these downsampling and decimation operations.

We thank the reviewer for noticing the lack of this important detail. We have used the standard implementation of the Python statistical package 'scipy' to decimate the signals, which includes an anti-aliasing filter. We have now added this information in the text:

[lines 319-320] "Thus, we decimated our signal from 48 kHz to 16 kHz, after applying an anti-aliasing filter."

[lines 347-348] "The audio signal was downsampled to 16 kHz, after applying an anti-aliasing filter, and we extracted the speech spectrogram using the Short-Time Fourier Transform."

- Line 258: Add a few sentences to explain the procedure. Also, the code could use further high-level explanations in the header.

We have added details to replace "Using a methodology similar to Gao et al.":

[lines 322-325] "Following the methodology described by Gao et al., the system first extracts acoustic features from the speech signal, including mel-frequency cepstral coefficients (MFCCs) and voicing information. These features are then processed through a pre-trained Long Short-Term Memory (LSTM) neural network that estimates the underlying articulatory movements. Using this methodology, ..."

- Line 272: I find it confusing that you explain three times the same preprocessing procedure for the neural data. Is there a reason? Am I missing something?

We apologize for the confusion. The preprocessing steps (re-referencing, frequency extraction, line noise removal) were the same for all three speech representation and therefore described in the 'Preprocessing' section prior to the representation pipelines, but the subsequent processing steps (including windowing) were different between the pipelines and therefore described in the individual sections. However, the size of the segmented windows and frameshifts were indeed shared between the articulatory and acoustic pipelines (but not the semantic one). To avoid unnecessary repetition, we removed the information about windowing from the individual sections and moved it to the 'Preprocessing' section, which we now re-named 'Neural data processing':

[lines 295-305] "We segmented the resulting signals in 50 ms windows with a 10 ms frameshift for the articulatory and acoustic representations to capture the complex dynamics of speech. Since neural activity precedes (in case of production) or succeeds (in case of perception) behaviour, we shifted the neural and audio data ranging between -200 and +200 ms in steps of 50 ms. These time-shifted neural signals were added as 'timeframe' features for the articulatory and acoustic data analyses, representing additional temporal context. The neural signal was segmented into larger windows for the semantic representation than for the other two representations, since semantic information is expected to have slower dynamics. This data was segmented in 200 ms windows with a 25 ms frameshift instead."

- Line 268: I am not sure I understand. You are preprocessing and aligning the stimulus feature and the neural signal. Then, you perform a linear regression to quantify the relationship between the two. You say that you are "predicting" articulatory features for each channel. Does that mean that you are "decoding" articulatory features from a single channel? Does that include any time-latencies, or is this model only measuring instantaneous relationships? i.e., what do the authors mean exactly by "each timeframe"?

We thank the reviewer for pointing out a confusion in the way the methodology is currently described. Initially, the pipeline was indeed built as a multivariate decoder, including the timeframes (temporal context) as additional features. However, since we later determined that evaluating the different timeframes to elucidate temporal dynamics would be interesting, we changed this 'decoder' to be univariate (per channel, per timeframe). This is essentially the same as directly correlating the neural signal with the audio or articulator, leading to the exact same results. This section is indeed still written as if it were a multivariate decoder 'predicting' or 'estimating' the audio/articulator, which is unnecessarily complicated. We have changed the text and associated code to the use of correlations instead.

Articulatory representation:

[lines 336-342] "We correlated the articulatory trajectories with the neural data from each channel and their time-shifted neural features (time-frames) using Pearson correlation to find the neural articulatory representation. This was calculated in 10 folds to match the 10-fold cross-validation in the semantic pipeline. The folds were averaged to obtain the final score per channel, time-frame and articulator. The results from the articulators were pooled together to form one articulatory representation (described further in 'Significant channels')."

Acoustic representation:

[lines 350-354] "We then correlated, also in 10 folds, each mel-bin (N=23) separately with each single channel and their time-shifted neural features (time-frames) using Pearson correlation. The folds and mel-bins were averaged to calculate the final score for each channel and time-frame."

- Data sharing: I gave the data on OSF a quick look. Is the neural data in there? The documentation is insufficient. There should be at least a document explaining what all the files represent.

We apologize if this is not clear enough, the code repository as an entry point to this dataset contains a description of all data files. The processed neural data is included in each representation folder and named `'*_features.npy'`.

Reviewer #2 (Remarks to the Author):

This study uses the sEEG system to investigate where articulatory, acoustic, and semantic representations are processed in the brain. The articulatory, acoustic, and semantic representations were estimated based on Acoustic-to-Articulatory Inversion Recurrent Neural Network, compressed speech spectrogram, and semantic embeddings from Word2Vec respectively. The study aimed to identify significant channels that contributed to processing of one of these representations or a combination of these representations. The study concluded that both hemispheres contributed to processing of all representations, however, there was greater involvement of the left (compared to the right hemisphere). The semantic representations were processed by widely distributed channels, which were distinct from the channels responsible for processing articulatory and acoustic representations. The conclusions provide converging evidence on localization of articulatory, acoustic, and semantic representations.

I have four major issues with the study at least in the form in which it is currently presented.

We thank the reviewer for their time invested in improving our manuscript and address each point individually.

1) In the abstract and throughout the introduction the readers are promised that this study will "limit potential con-founders" or help "acquire crucial complementary information for speech BCI's". I do not see how the current results bring us closer to be able to use BCI's or how this study helps to "open the path for speech neuroprostheses". Perhaps a more in-depth discussion of how the results of this study helps with this aim could be added to the discussion. Additionally, I wonder to what extend mapping of

representations for individuals with unimpaired speech will help decode speech in individuals that experience speech deficits due to brain impairments (e.g. stroke, TBI)?

The reviewer makes valid points. With limiting potential con-founders we were referring to variations in participant demographics, environmental conditions, psychological conditions (i.e., mood) and experimental setups when analyzing data between conditions as is often done (as compared to deriving these features from the same data). However, we indeed don't address this further in the introduction. In combination with the other remarks about the 'promises' in the abstract, along with those from reviewer #1, and the guideline from the journal to shorten the abstract, we have decided to remove this entire sentence from the abstract.

We have also toned down the statement in the conclusion:

[lines 254-255] "In combination, our results shed further light on the distributed and widespread networks involved in the speech production process, **which might be utilized in the development of** speech neuroprostheses based on differential representations of speech."

Indeed, there is still a large gap between these findings from individuals with unimpaired speech to those with brain damage. Current state-of-the-art speech BCIs focus on the motor cortex in patients with speech impairments due to ALS (Willet et al., 2023, Nature: <https://doi.org/10.1038/s41586-023-06377-x>) or a brain-stem stroke (Metzger et al., 2023, Nature: <https://doi.org/10.1038/s41586-023-06443-4>). This approach may indeed not work when the motor cortex is damaged due to a stroke or TBI. Our research does not directly improve speech BCIs for these individuals. However, it may inform the development by showing which (alternative) anatomical regions are involved generally and how other representations of speech may be utilized. For example, when a person cannot produce an articulatory representation due to Broca's aphasia, they might benefit from a BCI that works through a semantic representation instead. We gave such an example in the introduction (line 51: "While these studies have decoded speech from motor or acoustic representations of speech, individuals with diseases that cause impairments in the creation of motor representations, such as with aphasia, or in the generation of audible speech, such as with locked-in syndrome, might benefit from the decoding of more semantic representations of speech."). This work could therefore act as a guide for further research (line 71: "We aimed to provide an overview of how multiple speech representations are neurally distributed throughout the brain"). Short discussions on how the results could inform speech neuroprosthesis development are woven into the general discussion (e.g., lines 205, 220, 223, and 241).

2) The way the methodology is currently described doesn't allow for replication. Crucial information on how the representations were calculated is difficult to find or missing. For example, you mention "21 parameters represent trajectories of different articulatory muscles". Are these parameters pre-specified by the pipeline, or were these selected for this experiment? It is mentioned on line 268 that "We trained a linear regression in a 10-fold cross-validation... The folds were averaged to obtain the final score per channel, timeframe and articulator". Where exactly can I find these results, because I did not see results per articulator in the paper. Could you elaborate on the use of semantic embeddings in the linear regression? In general, the whole methodology should be made more detailed.

The methodology could indeed benefit from more details, we have added details in general and more specific points are discussed below.

The parameters are indeed pre-specified by the AAI pipeline and we did not make any particular selection for the current study. We have tried to clarify this in the text and added details:

[lines 322-328] **"Following the methodology described by Gao et al., the system first extracts acoustic features from the speech signal, including mel-frequency cepstral coefficients (MFCCs) and voicing information. These features are then processed through a pre-trained Long Short-Term Memory (LSTM) neural network that estimates the underlying articulatory movements. Using this methodology, we extracted a fixed set of 30 time-series related to the articulation process for each utterance. These 30 parameters are integral to the model, it includes eight parameters that are fixed values for a potential subsequent articulatory-to-audio conversion, and one that is a binary estimation of speech versus silence ('subglottal pressure')."**

Regarding the results per articulator (line 268), we described in the later 'Significance channels' section (lines 302-304) that the results were combined: "Any channel that had a correlation above the articulatory significance threshold in any of the articulators and timeframes was deemed significant for the articulatory representation. Thus, the results from individual articulators were pooled together to form one articulatory representation". This was done because the procedure is in this way more similar to that of the semantic pipeline and it would facilitate the presentation of the results. We have now added the information about 'pooling' within the articulatory section as well to avoid the confusion:

[lines 341-342] "The results from the articulators were pooled together to form one articulatory representation (described further in 'Significant channels')."

We have also added an elaboration about the use of semantic embeddings for the linear regression in the semantic representation pipeline, along with other details:

[lines 356-371] **"As opposed to the other two representations, the semantic representation was analyzed on a trial-based level.** The whole speech trial (2 seconds) was assigned to the cued word. Semantic embeddings (160 dimensional **numerical representations of words capturing their semantic meaning**) were extracted for each Dutch word using a pre-trained `\textit{word2vec}` model `\cite{tulkens2016evaluating}`. **The word2vec model was chosen for its simplicity and suitability for single-word representations, avoiding the contextual complexity of more modern models that are unnecessary for single words.** The embeddings were available for 85 of the 100 words, thus the 15 word trials without embeddings were discarded for each participant. As opposed to the previous two representations, we chose **to use** an encoding model for the semantic representation. A linear regression was trained to predict the neural timeseries for each recorded channel individually from the **semantic** embeddings, in a 10-fold cross-validation (Fig. 1D). **Linear regression was chosen for its interpretability and suitability for the relatively small number of word trials, allowing us to capture the linear relationship between the semantic features and neural activity.** Performance was evaluated using the Pearson correlation between the original and the predicted timeseries per window (**time-point**) across word trials for each single channel. **This procedure evaluates temporal variability in the response to spoken words rather than measuring how well the average response is captured."**

3) I am wondering about the appropriateness of the statistical tests used. The later sections of the results (anatomical contributions to speech representations, temporal dynamics of speech representations), have a "descriptive character" and I am missing some statistics based on which the conclusions can be drawn. This is especially true for

statements like: “the semantic representation appears to be stronger in the right hemisphere in these primary regions specifically”. I am also wondering whether drawing the main conclusions based on t-tests is the most appropriate. Does other literature in this field use this approach?

It is indeed true that the results are presented in a descriptive manner. We apologize if this style doesn't match the expectations of the reviewer, however, we do believe this style is common in our subfield (e.g., Nature: Willet et al., 2023, <https://doi.org/10.1038/s41586-023-06377-x>; also in this journal (Communications Biology): Angrick et al., 2021, <https://doi.org/10.1038/s42003-021-02578-0>). We additionally believe this facilitates reading comprehension and would therefore like to keep this style of writing. The descriptive style also comes from the fact that we present the results as a total number or percentage of significant channels (of which the relative difference between the representations is of interest). We chose this metric to allow a fair comparison between the representations. The strength of the correlations would not be fair considering the different pipelines and different chance levels. The significance is, however, statistically determined through permutation tests. Presenting the results across individuals, rather than on a whole-brain level, would also not be fair considering the large differences in electrode coverage between participants. Additionally, the 'performance' within individuals were not of interest in this work, but the whole-brain dynamics are. Since the number of significant channels are single numbers, we could not apply additional testing to support these descriptions. However, we were able to quantify the spatial distribution of the representations through the distribution metric, on which we were able to run statistics. The main conclusions were drawn from these statistics, along with descriptions regarding specific anatomical regions and temporal results.

It is common in our field to run pairwise comparisons without a main effect analysis (e.g., the previously mentioned Angrick et al., 2021 and Metzger et al., 2023). The Mann-Whitney U tests were used in these papers due to non-parametric data. On closer inspection, two of our six groups were non-parametric. Therefore, we decided to change the t-tests to Mann-Whitney U tests as well. For sake of completeness, we also performed the Kruskal-Wallis Test for the results per hemisphere to investigate the main effects. These were significant, which would allow the pairwise comparisons. The outcome of these tests remained the same as for the previous results using t-tests. We have adjusted the text and added details in the results and methods sections regarding the distribution metric.

Results:

[lines 100-124] **“To quantify the spatial distribution of the results (Fig. 2D), we calculated distances between the location of significant channels and their centroid within representation and hemispheres. A Kruskal-Wallis test was conducted to compare the distances between the articulatory, acoustic, and semantic representations in the left and right hemispheres. There was a significant main effect between the groups in both the left ($H(2) = 48.42$, $p < 0.001$) and the right ($H(2) = 31.16$, $p < 0.001$) hemisphere. Post-hoc two-sided Mann-Whitney U tests revealed significant differences between articulatory and semantic distances (Left: $U = 1095$, $p < 0.001$; Right: $U = 264$, $p < 0.001$) and between acoustic and semantic distances (Left: $U = 1108$, $p < 0.001$; Right: $U = 323$, $p < 0.001$), but not between articulatory and acoustic distances (Left: $U = 3189$, $p = 1.0$; Right: $U = 487$, $p = 1.0$). Another three two-sided Mann-Whitney U tests were performed to compare the distances between hemispheres within the representations. These tests revealed significantly greater distances in the left hemisphere for all three representations (Articulatory: $U = 1889$, $p < 0.001$;**

Acoustic: $U = 2019$, $p < 0.001$; Semantic: $U = 2139$, $p = 0.02$). All post-hoc tests were strictly corrected using the Bonferroni correction for multiple comparisons ($N = 9$).

These results indicate that the semantic representation is more widely distributed across the brain than the other two representations, in both hemispheres. Importantly, this holds true within individuals (14/15 participants). Only for one individual the articulatory-selective channels were spread farther than the semantic ones (Fig. S2). **The results also indicate that** all representations are more widely distributed across the left than across the right hemisphere. This difference could potentially be due to sampling, since the electrode coverage is not exactly mirrored between the hemispheres (along with more contacts in the left hemisphere overall). Therefore, we additionally calculated the distance between all recorded channels and their centroid in each hemisphere. This **analysis did not reveal a statistically significant difference between the distances of** the left (38.47 ± 16.69) and the right (37.25 ± 17.57) hemisphere overall (**$U = 350849$, $p = 0.06$**). The difference in the distribution of the representations between hemispheres is therefore not likely due to the electrode coverage alone. Together, these findings indicate distinct and widespread processing of the meaning of words."

Methods:

[lines 394-399] **"Kruskal-Wallis and post-hoc two-sided Mann-Whitney U tests** (Bonferroni-corrected for nine repeated tests) were used to calculate the statistical difference in distribution (i.e., distances) between the three representations within the hemispheres and between the left and right hemisphere within representations. **The Euclidean distances were additionally calculated between all recorded channels and their centroid within each hemisphere, to rule out differences due to sampling alone, and tested using a two-sided Mann-Whitney U test. These non-parametric tests were chosen as they do not assume normally distributed data."**

With regards to the statements in the 'Anatomical contributions to speech representations', these are indeed qualitative and not statistical comparisons due to the nature of the results we present (the number of significant channels). We have clarified it in the manuscript:

[lines 128-129] "Here, we provide a qualitative examination of the data to describe general patterns within and across ROIs."

We have also clarified the use of qualitative comparisons in the 'Temporal dynamics' result section:

[lines 139-141] "To explore the brain-wide temporal dynamics, including channels that are not represented in ROIs, we evaluated significant channels per time-point or time-frame **for a qualitative temporal analysis** and grouped these results into three larger time segments **for a qualitative spatial analysis**."

4) I find that one of your important findings is not discussed in relation to other literature anywhere in the discussion (see p. 4: "The temporal dynamic results indicate that the majority of articulatory- and/or acoustic-selective channels are active well before speech onset... This could reflect the communication between different areas for speech preparation")

Indeed, we have added relevant literature, including a very recent preprint on the coordination of regions in speech comprehension (Gwilliams et al., 2024:

<https://doi.org/10.1101/2024.04.19.590280>):

[lines 173-174] "This finding is especially pronounced in the left hemisphere, even in auditory regions. This could reflect the coordination between different areas for speech preparation^{10,11,43}, which has similarly been found for speech comprehension⁴⁴."

Minor Issues:

1) I worry that the readers might misinterpret the following sentence: "there appears to be a relatively stronger tuning for articulatory features in the left hemisphere versus a relatively stronger tuning for acoustics in the right hemisphere (Fig. 2C)". Some might read this as "there is stronger tuning for acoustics in the right hemisphere in general". Could you perhaps rephrase?

That could indeed be a possible misinterpretation. We changed the text to make this distinction clearer:

[lines 95-97] "..., there appears to be a stronger tuning for articulatory features relative to acoustic features in the left hemisphere and the opposite, a stronger tuning for acoustic features relative to articulatory features, in the right hemisphere."

2) I am not in agreement with the interpretation of this null result "This revealed no difference between the left (38.47 ± 16.69) and the right (37.25 ± 17.57) hemisphere overall ($t(1645) = 1.44$, $p = 0.15$)". Bayesian statistics would have to be used for the claim of "no difference".

We agree that a claim of "no difference" requires stronger statistical support. To avoid overinterpretation of this null result, we have rephrased the text and changed the t-test to a Mann-Whitney U test:

[lines 118-120] "This **analysis did not reveal a statistically significant** difference between the distances of the left (38.47 ± 16.69) and the right (37.25 ± 17.57) hemisphere overall (**$U = 350849$, $p = 0.06$**)."

3) The section titles "Anatomical contributions to speech representations" are too general, as this study only discusses distinct aspects of speech representations.

This study is indeed focused on three specific aspects of speech representations. Since this is made clear throughout the text and figures, we trust that the reader will understand that 'speech representations' in this subtitle, similar to the other subtitles, refers to the three speech representations that are included and discussed in this section and this work in general.

4) When grouping results into 3 time segments. Why did you opt for segments of different lengths?

We are sorry if the grouping of time segments was not clear. The 3 time segments in the articulatory/acoustic features are all 100ms in length and the 3 time segments are all 625ms in length for the semantic features. The difference between the representations stems from the differences between the pipelines (continuous correlations vs trial-based encoding), and they have slightly different meanings. We highlighted this difference in the 'Temporal dynamics' methods section. Since the results are presented before the methods, we have now elaborated the explanation in the results section as well:

[lines 139-143] "To explore the brain-wide temporal dynamics, including channels that are not represented in ROIs, we **evaluated significant channels per time-point or time-frame for a qualitative temporal analysis** and grouped these results into three larger time segments **of equal length** for a qualitative spatial analysis. The time-frames represent **neural features that were temporally shifted to be** prior to speech alignment (-200 until -50 ms), surrounding speech alignment (-50 until +50 ms) and after speech alignment (+50 until +200 ms) for the articulatory and acoustic representations."

5) The time windows for the semantic representations are not clear from the text (p. 3).

We hope that the time windows (now called time-points) are now clearer with specifying semantic 'time-points' and articulatory/acoustic 'time-frames' in this section:

[lines 139-144] "To explore the brain-wide temporal dynamics, including channels that are not represented in ROIs, we evaluated significant channels per **time-point or time-frame** for a qualitative temporal analysis and grouped these results into three larger time segments of equal length for a qualitative spatial analysis. The **time-frames represent** neural features that were temporally shifted to be prior to speech alignment (-200 until -50 ms), surrounding speech alignment (-50 until +50 ms) and after speech alignment (+50 until +200 ms) for the articulatory and acoustic representations. For the semantic representation, the **two-second speech trial time-points** were equally divided."

6) I would like figure S1 to be a lot clearer. Could you perhaps include one figure per participant, so the readers can get a better idea about electrode placements?

While we think it is important to show the electrode placements from individuals in the averaged MNI brain space as this is the space the results are presented in, we have additionally added an axial view of the electrode placement in each individual's native space for clarification:

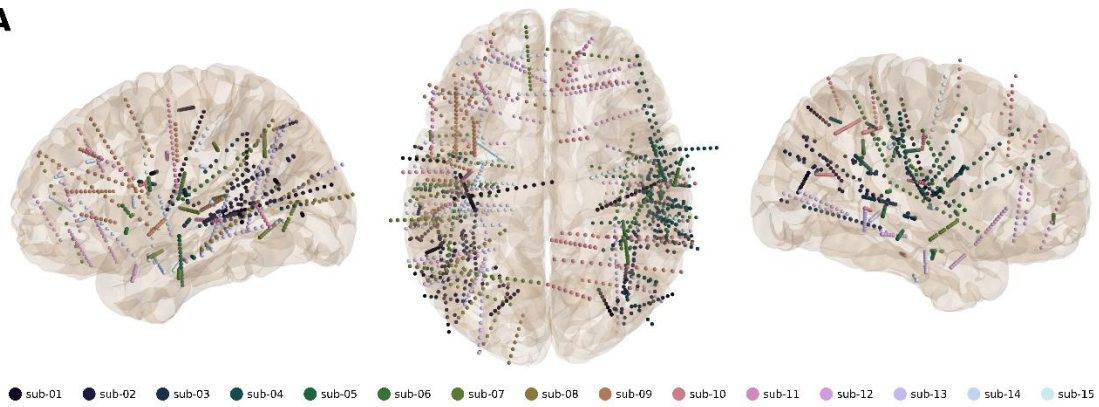
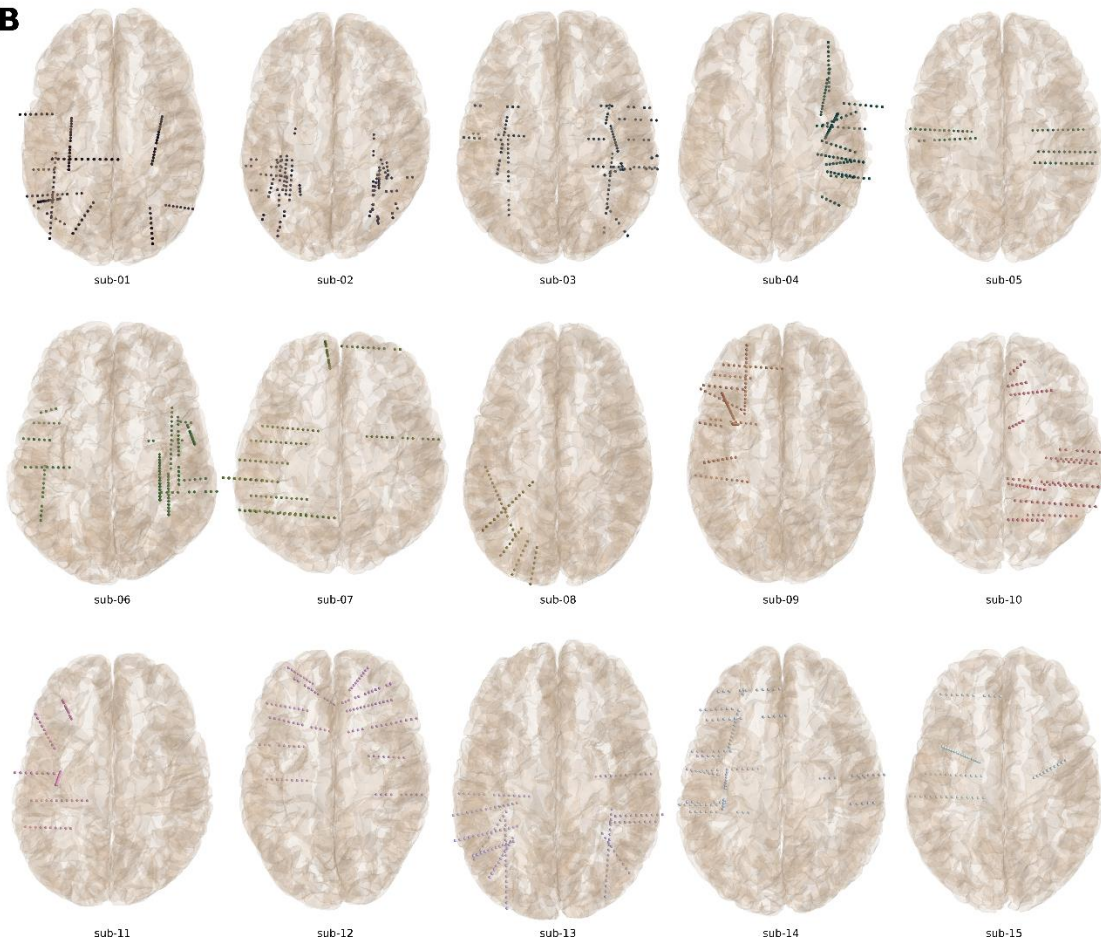
A**B**

Figure S1. Electrode locations. **A)** Electrode locations within the MNI averaged brain using MNI coordinates from individual participants. These locations correspond to those in the main figures. Colors represent the individual participants. **B)** Electrode locations within an axial view of each individual participant's native brain.

REVIEWERS' COMMENTS:

Reviewer #2 (Remarks to the Author):

The authors have addressed the issues I raised carefully. They provided very thorough explanations and justifications of their chosen analyses and writing style. The methods and results are a lot clearer in the current version. There are a couple of minor clarifications that could benefit the reader in the beginning of the result section:

1. On line 89: "For all three speech representations (articulatory, acoustic and semantic), there are significant channels present in both hemispheres (Fig. 2A)." It may be helpful to briefly reiterate how the significant channels were calculated to aid reader comprehension.

We have moved the methods summary from the last paragraph of the introduction to the beginning of the results section and added the following information:

Significant channels were identified by estimating a distribution of chance correlation coefficients through random permutations and circular shifts, setting a strict significance threshold ($\alpha = 0.05$) using max- t correction, and determining channels as significant if their correlations exceeded this threshold in any time-frame for each of the representations separately.

2. On line 113: "Only for one individual the articulatory-selective channels were spread farther than the semantic ones (Fig. S2)." Figure S2 shows the results separately per hemisphere (which is informative for the readers), however, it is not immediately clear that statement above refers to averages across both hemispheres.

The reviewer is correct in that the statement did not refer to averages across hemispheres, but to the findings within the left hemisphere in particular. We've clarified the text in the following way:

Importantly, this holds true within individuals (8/10 participants with significant channels in all three representations). The articulatory-selective channels were spread farther than the semantic ones for only one individual and the acoustic-selective channels for two individuals, both within the left hemisphere (Supplementary Fig. S2).

The current study provides valuable contribution to studying neural dynamics of various stages in the speech production process. I look forward to seeing the manuscript published.

Reviewer #3 (Remarks to the Author):

This revision has addressed my concerns. Well done on an excellent study.

I have a couple of minor comments (the current version can be published as is, in my view).

- Lines 82-85: The authors use a valid method, and the explanation is sufficient. I find it somewhat unconvincing in that there is no explicit mention to why the same things could not be done with an encoding models (I can imagine the motivation, but it's not sufficiently explicit in the text). Nonetheless, I find the explanation sufficient for a generalist reader.

Thanks to the previous round of reviews, we noticed that the use of decoding (or encoding) models was misleading as we're dealing with single features (signals per

channel, per timeframe) for the articulatory and acoustic representations. The descriptions and analyses were therefore changed to correlations. However, an intermediate encoding step was still necessary for the semantics due to the dimensionality of the embeddings. A mixture of decoding and encoding models would indeed have warranted a discussion on this limitation, however, we believe the change in analysis has rendered such a discussion irrelevant.

- Regarding the title (my final major comment from the previous review). We agree to disagree. I still think that the results of this paper are primarily on the spatial dimension. So, a title reflecting that would be more appropriate in my view. It's not a deal-breaker for me, so I'll leave this to the editor.

We greatly appreciate the trust from both the reviewer and editor and, after careful consideration, have decided to keep the original title.