nature biomedical engineering

PEER-REVIEW INFORMATION

**A Spanish–English speech neuroprosthesis driven by multilingual cortical articulatory representations**

Corresponding author: Edward Chang

## Editorial note

This document includes relevant written communications between the manuscript's corresponding author and the editor and reviewers of the manuscript during peer review. It includes decision letters relaying any editorial points and peer-review reports, and the authors' replies to these (under 'Rebuttal' headings). The editorial decisions are signed by the manuscript's handling editor, yet the editorial team and ultimately the journal's Chief Editor share responsibility for all decisions.

Any relevant documents attached to the decision letters are referred to as **Appendix #**, and can be found appended to this document. Any information deemed confidential has been redacted or removed. Earlier versions of the manuscript are not published, yet the originally submitted version may be available as a preprint. Because of editorial edits and changes during peer review, the published title of the paper and the title mentioned in below correspondence may differ.

## Correspondence

Wed 11 Oct 2023
**Decision on Article nBME-23-1527**

Dear Dr Chang,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "A bilingual speech neuroprosthesis" and forgive us for the delay in making a decision, we were waiting for the last report that took longer to be delivered. The manuscript has been seen by 3 experts, whose reports you will find at the end of this message. You will see that the reviewers appreciate the work, and that they raise a number of technical criticisms that we hope you will be able to address. In particular, we would expect that a revised version of the manuscript provides:

* Extended discussion of the limitation of a small sample size (n =1) with emphasis on how it impacts the generalization of the findings to application in other patients;

* Thorough methodological reporting;

* Additional controls, as per the relevant comments of Reviewers #1 and #3.

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We hope that you will be able to resubmit the manuscript within 20 weeks from the receipt of this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*.

We hope that you will find the referee reports helpful when revising the work, which we look forward to receive. Please do not hesitate to contact me should you have any questions.

Best wishes,

Valeria

__
Dr Valeria Caprettini
Associate Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

This study describes a method to decode bilingual speech production with invasive electrophysiology and language model of English and Spanish. The participant is a native spanish speaker (L1) who learned English (L2) later in life and suffered from brain stem stroke that made him unable to produce intelligible speech, just grunts and moans when attempting speech.

The authors have done an excellent job describing the details of their approach. The manuscript is written to be comprehensible to large audience.

A major claim is that despite later acquisition of a second language (L2) later in life, there was no dedicated cortical regions or patterns of neural activity that are unique to L2 (English) or L1 (Spanish) speech attempts. The authors conclude that their findings align with theories that the brain fits L2 into the articulatory framework of L1, suggesting a shared cortical representation between languages.

The impact of the study could significantly reduce the required training times for multilingual BCI use.

Main points:
1- The sample size of the linguistic context model used is small (e.g. syllable paired words Table S8). The authors acknowledge that needing to build linguistic context to decode the intended language is one limitation of this study.

2- While they demonstrated a strong shared articulatory representation between English and Spanish, it is unclear whether this finding will generalize across subjects who learned L2 at different times. Authors need to state how long the subject started learning L2 after L1 and the extent to which this could have contributed to this shared representations. One could argue that the longer the time between L1 and L2 learning, the

less these shared representations might be, as evidenced by Spanish speakers who speak fluent English when they learn it at childhood versus those who learn English in adulthood. As this is a single subject study, the question of shared representation will likely remain open.

Detailed comments:
1. Line 512: The 'Data included in this paper, is from a subsequent version of the task where the participant was allowed to attempt audible grunts of the target word (overt). Have the authors attempted to examine neural features in trials in which the subject was instructed to attempt any word not present in the dataset? This is important since the shared features could just be the encoding of the grunts rather than the specifics of the word being instructed. Also it is unclear how different the neural features were when the participant was asked not to vocalize (mimed) versus when he was.

2. Line 527: It is stated that presentation of the target word on the screen and the go-cue were labeled as 'speech preparation' and time points between the go-cue and 2 seconds after the go-cue were labeled 'speech'. Since attempted words could span different lengths, possible 'rest' labels could be included in a 'speech' window label. Please clarify how this affected the results.
The decoding rule relies primarily on the fixed 3.5 sec used to extract neural features. It is unlikely that attempted speech will always abide by this fixed length; some could be shorter others could be longer. Have the authors tried different window lengths to evaluate decoding phrases in particular (Fig S1)?

3. Have the authors compared neural activity periods where the subject silently read or listened to the words to periods of attempted speech? It would be important to fully understand whether the attempted speech shared anything with reading or listening to that speech in order to ascertain that the neural representations do indeed reflect motor production of speech.

4. How error in the word error rate (WER) metric was split between neural only versus LLM only, in case of isolated versus phrase contexts?

5. Fig2a shows significant differences between English and Spanish word classification rates. Can authors comment on why this is the case.

6. Fig2: Clarify the extent of decoder training between sessions separated by 30 days breaks

7. AAC speed was only based on 8 sentences (4 English and 4 Spanish). This estimate of 6.59 words per minute might be largely biased. Can authors comment on how this could affect significance in Fig1?

8. It is stated the HFG 70-150 and LFS 0.3-100 Hz were acquired and used as features to classifiers.Given the band overlap, how did this redundancy in the 70-100 Hz band affect classifications? Is there a reason for extending the LFS above 70 Hz?

9. Why did they have to train the classifiers on silent isolated word attempts before they started to retrain on the attempted grunts/moans? How was the detector threshold developed and fine tuned in such case?

Reviewer #2 (Report for the authors (Required)):

The study by Silva et al. from the Chang lab is an extraordinarily clear and clean demonstration, both of a brain function and of a technical small masterpiece: The real-time, soft- and hardware-aided decoding of language and language content from sensori-motor, articulatory (i.e., neural) representations in the human brain, which allows the authors to promote this as a demonstrator of a bilingual speech neuroprothesis. For once, this is hardly an exaggeration.

The manuscript is technically advanced but for most uses established neural encoding/decoding routines long established by the lab and other ECoG/invasive-ephys researchers, combined with the power of large language networks (here, GPT-2 and word2vec if I am not mistaken). The strength of this manuscript and its novel aspect thus clearly lies in now approaching the question of bilingual speech representation and the feasibility of bilingual neuroprotheses.

I consider this m/s as presented a fortunate instance of a two-way street between neuroscience and

neuroengineering. A long-standing question on bilingual representations gets tackled and in part answered by the data presented (see eg. Fig. 3), but doing so results from a more engineering-geared question asking why the bilingual decoder works in the first place.

I have been pondering the major or minor concerns one could or should additionally bring forward against the manuscript in its current form. However, due to the fact that this is by all intent a proof-of-concept report and not an inference-on-population study, the criteria are clearly different. For a single, highly-specialised patient this is by all means an impressive, and methodologically clear report with neuroscientific and neuroengineering value.
While different labs might have taken different analytical choices, the report as is is transparent and sound enough to clearly allow the reader to judge for themselves. Minor quibbles on why the outdated/underpowered Wilcoxon/Mann -Whitney (instead of permutation tests) keeps being should thus not be a point of contention here.
I would recommend that the Discussion section is amended with a paragraph on potential roads to application in other patients or its applicability to non-invasive neuroscience (citing e.g. Correira et al., 2014, https://www.jneurosci.org/content/34/1/332).

Reviewer #3 (Report for the authors (Required)):

Silva et al. present a study that extends a prior proof of concept brain to text system (Moses et al., NEJM, 2021) from English only to a bilingual (English and Spanish) system. The core demonstration and additional analyses will be valuable to a burgeoning field and provides forward vision towards the broader translation and use of this technology. Their core demonstration is that the same prosthesis can be utilized for the L1 (Spanish) and L2 (English) languages of their participant and that transfer learning approach (commonly utilized by the broader machine learning community) can be successfully employed to more rapidly train and adapt the neural prosthesis to a second language. Their findings, consistent with our expectations from prior literature, suggest that the overlap in neural activity present in both languages that enables their technical demonstrations is the result of a shared articulatory representation for both languages.

Although there is only one participant in this study, given the challenging nature of this work and the advance in knowledge even an N of 1 provides in this field, it should be strongly considered for publication. N of 1 publications in this field are increasingly common, for example, (Moses et al., NEJM, 2021), (Willet et al., Nature, 2021), (Willet et al., Nature, 2023), (Metzger et al., Nature, 2023).

I have a few major and minor concerns about the work that, I believe, should be addressed prior to full consideration for publication:

1) Throughout the manuscript the term “real time” is used. For a broad bioengineering audience, the use of this term could lead to confusion. "Real time" for many engineered systems implies microseconds or milliseconds of latency. It can also be interpreted as a system with output delay (latency) that is low or negligible for the given application. In the case of a speech and language system there are likely a few key numbers that could be considered — for example if this was a speech synthesis system a delay of 10s of milliseconds could be argued to be real-time, as such a delay introduced for able bodied speakers does not result in appreciable speech deficits. Since this is a text based output system, I think it would be reasonable to look at the median turn taking time between two healthy individuals, which is on the order of 100s of milliseconds. However, if I understand your system architecture correctly (and from watching the supplementary videos), the latency is multiple seconds. Thus, I think real time is not an appropriate description and could lead to misunderstanding by many readers. It might be better to describe this system as operating “online.” If you do use “real time” it should always be qualified with the effective latency between speech intent and output to screen.

2) Related to the previous concern, the schematic of the decoder in Figure 1a may mislead some readers. The use of 3.5 seconds of data along with a bi-direction RNN could be better highlighted. The current figure could be misinterpreted to be a uni-directional RNN with shorter input segments. I’d suggest augmenting Figure 1a to more clearly show the flow of data / relative timing. I’d also suggest editing the Supplementary Methods section S1 and adding a Supplementary Figure that more fully describes the algorithms/architecture and temporal flow of data.

3) I think the definition of “stability” needs to be better qualified. The comparison to intracortical techniques and results are a bit confusing and misleading. In this case, the authors are highlighting system performance without daily recalibration. However, for the field it is important to separate the stability of neural features (i.e., stable relationship between neural modulation and intended language output) vs. stability in the information content in the neural features (i.e., relationship between neural modulation and intended language output is not stable, but overall information content is stable). For this work, it is interesting to see that performance does not trend downwards when the same decoder is used from day to day, but the word “stable” is misleading, as the graphs show quite a bit of variability. There is about a 20% difference between the top performing and lowest performing day — I don’t think it’s reasonable to consider decoding performance of 50% - 70% to be “stable.” If we compare this variability to the recent Willet et al, Nature, 2023 study, one could argue that their intracortical result is far more stable — Fig 2b in that paper shows a much greater stability during online control sessions (~5% difference between lowest and highest performing days) for both a similar complexity task (50 word vocabulary) and higher complexity task (125,000 word vocabulary). Of course, an important caveat in that study is that they are retraining the decoder everyday, whereas in this study the decoder was fixed. A valuable course of action would be to compare performance when retraining each day to performance when holding the model fixed. Since all of the characterizations Figure 2 of this manuscript are offline, these analyses should be straightforward to conduct. Then the results should be presented with proper qualification of the word "stable" under these two very important contexts.

4) Is chance performance calculated with the language model in place? I believe this is the case based upon line 186 in the manuscript. Given that the chance calculation is based upon a shuffle and the core metric of the work is a word error rate, I would like to see a “neural activity only” equivalent for testing chance. This would allow the reader to better disambiguate the role of real-time neural observations vs. integration across time through application of a language model. An understanding of this tradeoff will be a critical design criteria for practical systems and for considering a tradeoff between system latency, vocabulary size, and performance. Why is chance error sometimes greater than 100%?

5) The transfer learning result is tantalizing. However, the workup of the result is quite limited and would be greatly strengthened by evaluating the shared aspects of the two languages that facilitate transfer learning — the results shown suggest that pre-training with either language results in equivalent reductions in required training set size. Training set size is a major consideration for the design of these systems — as shown in figure 3, without pretraining 6+ hours of subject specific training data are required to achieve performance that approaches the asymptote. How many hours of data were used for pre-training?
To better understand what shared aspects are at play, I would suggest that the authors examine the role of phonemic overlap / articulation feature overlap between the pre-train dataset and the target dataset for transfer. Examination of the impact on transfer can be conducted within language or across languages. Overlap can be controlled by sub-selecting phonemic or articulation features to alter the distributions shown in Supplementary Figure 3. This sub-selection could be conducted by design or randomly and the degree of overlap could be evaluated by measuring the KL divergence between the distributions of phonemic / articulation featured in the two datasets. Such a workup could provide better insight into the bilingual results of these paper and could also be helpful when considering transfer to other languages (and could help add to the wonderful discussion point when considering Mandarin). Furthermore, such a workup could also inform the construction of effective and efficient training sets for single language speech prostheses.

6) The introduction is missing references to similar studies that employ sEEG and intracortical electrodes. These should be included where appropriate in the introduction. Further, the implications for all implanted speech prosthesis should be discussed. Many of the findings in this work likely have implications for recent intracortical electrode based systems, as the system designs have many parallels, but also some unique aspects (e.g., lower effective latency in intracortical demonstrations / orders of magnitude larger effective vocabulary). This is also an opportunity to discuss some of the important subtleties that are critical design tradeoffs for these systems: achieved rate (which may also tie into natural speaking rates of the user and effective vocabulary size / limits to output flexibility caused by language model priors).

Minor comments and concerns:

1) Line 32, “Together, these findings establish a multilingual cortical articulatory representation that persists 33 after paralysis and enables flexible decoding of multiple languages.” — this is a proof of concept, N of 1, does not “establish”

2) Line 65 -> words -> word

3) Incorrect figure references? e.g. Fig. 2C on line 271?

4) “Confusability” should be defined / described in text

5) The manuscript switches from % error to % accuracy. This is confusing, why not stick to one of the two metrics?

6) Lines 217-218 — The phrasing of this paragraph is confusing. I believe the “overall” condition is when the language is not conditioned during decoding?

Mon 05 Feb 2024
**Decision on Article NBME-23-1527A**

Dear Dr Chang,

Thank you for your revised manuscript, "A bilingual speech neuroprosthesis". Having consulted with the original reviewers (whose comments you will find at the end of this message), I am pleased to write that we shall be happy to publish the manuscript in *Nature Biomedical Engineering*.

We will be performing detailed checks on your manuscript, and in due course will send you a checklist detailing our editorial and formatting requirements. You will need to follow these instructions before you upload the final manuscript files.

Please do not hesitate to contact me if you have any questions.

Best wishes,

Valeria

__
Dr Valeria Caprettini
Associate Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

I thank the authors for being responsive to my comments. The revised m/s version was substantially improved.

Reviewer #2 (Report for the authors (Required)):

I am content with the additional paragraphs and revisions added in response to my comments (and I also found the responses to other reviewers' queries quite on target and and to improve the m/s overall).
I have no concern over this being published as is.

Reviewer #3 (Report for the authors (Required)):

I appreciate the care with which the authors addressed reviewer concerns and for the additional experiments and analyses employed to answer our queries. The authors have satisfied my concerns and I believe the manuscript is ready for publication. I only have a couple of minor comments.

(1) Figures 2d, 2f, and 3b are each missing a color bar / legend. These would help the reader interpret each figure.

(2) I believe there is an error on lines 296 and 299, calling out Fig. 1g instead of Fig. 2g. I'd also suggest ordering the 3 numbered items to match their order on the corresponding plot.

**Rebuttal 1**

Below, we individually address the comments that we received. For each comment, we include:

- ***The original comment***
- Our response to the comment and a description of how the comment was addressed in the revised manuscript (in normal font)
- Any new figures that accompany our response to the comment
- The modified text in the manuscript (underlined; only included for some responses)

## Reviewer 1

1) **This study describes a method to decode bilingual speech production with invasive electrophysiology and language model of English and Spanish. The participant is a native spanish speaker (L1) who learned English (L2) later in life and suffered from brain stem stroke that made him unable to produce intelligible speech, just grunts and moans when attempting speech.**

   **The authors have done an excellent job describing the details of their approach. The manuscript is written to be comprehensible to large audience.**

   **A major claim is that despite later acquisition of a second language (L2) later in life, there was no dedicated cortical regions or patterns of neural activity that are unique to L2 (English) or L1 (Spanish) speech attempts. The authors conclude that their findings align with theories that the brain fits L2 into the articulatory framework of L1, suggesting a shared cortical representation between languages.**

   **The impact of the study could significantly reduce the required training times for multilingual BCI use.**

   We thank the reviewer for their positive and encouraging summary of our work and their detailed suggestions that we agree would strengthen the manuscript.

2) **The sample size of the linguistic context model used is small (e.g. syllable paired words Table S8). The authors acknowledge that needing to build linguistic context to decode the intended language is one limitation of this study.**

   We appreciate the reviewer's thorough reading of these points, and agree that currently linguistic context is needed to decode the intended language, which is a limitation of the system that we acknowledge and discuss.

3) **While they demonstrated a strong shared articulatory representation between English and Spanish, it is unclear whether this finding will generalize across subjects who learned L2 at different times. Authors need to state how long the**

**subject started learning L2 after L1 and the extent to which this could have contributed to this shared representations. One could argue that the longer the time between L1 and L2 learning, the less these shared representations might be, as evidenced by Spanish speakers who speak fluent English when they learn it at childhood versus those who learn English in adulthood. As this is a single subject study, the question of shared representation will likely remain open.**

We thank the reviewer for raising this point. The reviewer is correct that this is an n=1 study and we did not explicitly test how the results presented here of a strong, shared articulatory representation will generalize from our participant to other subjects who learned L2 at different times. To this end, we have added an expanded discussion paragraph where we discuss this limitation and also discuss how these findings may generalize to other participants and other formers of neural recording technology.

We have also clarified that our participant learned English (L2) later in adult life, reaching fluency at age 30 after his brainstem stroke. We agree with the reviewer that, in the literature, a predominant theory is that late acquisition of L2, similar to our participant, leads to a *lesser*, rather than larger, degree of shared representations [DeLuca et al. 2019, *PNAS*; Del Maschio et al. 2019, *Handbook of Multilingualism*; Shimada et al. 2105, *Neuroscience*; Cao et al. 2013, *J. Cogn. Neurosci*]. Thus, we feel that the results in this paper are even stronger evidence for shared articulatory representations since the reported subject learned L2 later in adult life, even after his brainstem stroke, adding a unique example to the bilingual speech production literature of strong shared representations despite late learning of L2.

Further, this is, to our knowledge, the first study of bilingual speech production with invasive electrophysiology. We feel that this offers a complementary perspective to the years of rigorous non-invasive bilingual speech-production research. By directly observing localized neuronal-population activity, we aimed to describe the content of the representations beyond focusing on magnitudes or other summary statistics. We feel that our decoding, confusability, syllable-decoding, and transfer learning analyses speak to this fact and complement the large body of non-invasive research. However, while we find it compelling that strong and shared articulatory representations underlie attempted bilingual speech in our participant, we agree that more work must be done to understand how this effect varies with proficiency, age-of-acquisition, and different languages. This study represents a proof-of-concept that these representations both exist and persist after paralysis and can enable bilingual decoding within a single system. We hope that these points are now better clarified with the additional discussion paragraph.

Although this is a single-participant study, the strength of shared cortical-articulatory representations is encouraging for generalizability to other patients. Notably, as this participant learned L2 (English) later in life, this may be even more encouraging for those who learn L2 early in life, which has been shown to lead to stronger shared representations 30,31,33,35. Future work, however, should directly investigate whether

this effect varies with L2 proficiency, age-of-acquisition, and articulatory similarity to L1. While this study leveraged invasive electrophysiology, the results also hold relevance for non-invasive BCIs. A potential advantage of non-invasive BCIs is the ability to sample a larger distribution of cortex involved in functions beyond articulation. Non-invasive BCIs may apply a similar framework, as applied here with articulatory features, for other shared language features, such as semantics 67,68, that could enable rapid generalizability of a BCI across languages. These studies may also find language-specific signals, which we did not observe in speech-motor cortex, that may exist in higher-order cortical regions 28,69,70. A complementary direction for invasive BCIs is to study whether language-specific signals exist at the level of single neurons.

The participant is a native Spanish speaker (L1) and learned English in adult life reaching fluency at age 30, after his brainstem stroke.

4) **Line 512: The 'Data included in this paper, is from a subsequent version of the task where the participant was allowed to attempt audible grunts of the target word (overt). Have the authors attempted to examine neural features in trials in which the subject was instructed to attempt any word not present in the dataset? This is important since the shared features could just be the encoding of the grunts rather than the specifics of the word being instructed. Also it is unclear how different the neural features were when the participant was asked not to vocalize (mimed) versus when he was.**

We thank the reviewer for this question and appreciate the opportunity to further detail the participant's behavior during the task. The participant has anarthria, diagnosed by a speech-language pathologist. His attempts to speak are unintelligible and are described by speech-language pathology as "grunts". Thus, by "grunts" we mean that the participant is attempting to speak to the best of his ability, albeit unintelligibly. We have added additional wording to the methods section to further clarify this point. A full SLP evaluation for this participant is also included in the supplementary information of Metzger et al. 2022, *Nat Comm.*

We also thank the reviewer for their suggestion to examine classification of neural features in trials in which the subject was instructed to attempt words not present in the dataset. We re-worked Figure 4 based on this suggestion and that of reviewer 3 (comment 7). We trained a model on a subset of words from the bilingual vocabulary (Fig. 4c). We then attempted to test the model on words that were not included in the training dataset (Fig. 4d, e). Interestingly, we found that we could not achieve above-chance performance on words that had different articulatory characteristics than the training set; however, strongly above chance performance was found on words that had similar articulatory characteristics to the training set. Overall, this provides evidence that grunts are not encoded, but instead the information encoded is the articulatory characteristics of the target words. With regard to the similarity of neural features

between overtly attempted speech (vocalized) and silently attempted speech (mimed), we have expanded on this in response 12 below. To summarize, these paradigms share an articulatory code that allows decoders to transfer across paradigms; however, there are some differences in the evoked activity patterns. A more detailed comparison of the two behavioral paradigms is provided in Metzger et al. 2022, *Nat Comm.*

5) **Line 527: It is stated that presentation of the target word on the screen and the go-cue were labeled as 'speech preparation' and time points between the go-cue and 2 seconds after the go-cue were labeled 'speech'. Since attempted words could span different lengths, possible 'rest' labels could be included in a 'speech' window label. Please clarify how this affected the results. The decoding rule relies primarily on the fixed 3.5 sec used to extract neural features. It is unlikely that attempted speech will always abide by this fixed length; some could be shorter others could be longer. Have the authors tried different window lengths to evaluate decoding phrases in particular (Fig S1)?**

We thank the reviewer for the opportunity to clarify these points and apologize for any confusion. Indeed there can be some variance in the length of each attempted utterance relative to the go-cue. Our choice of using 2 seconds after the go-cue as our cut-off for the attempt was roughly determined by looking at the average neural activity during speech attempts and visually determining when that activity was above baseline (on average). Undoubtedly, for some trials, this 2 second "offset" is actually before the end of the utterance. We treat the time points that are after this cutoff, but before the end of the trial, as ambiguous–the participant could either be finishing the attempt or could be back at silent rest–and do not include these timepoints in the training procedure. Because the speech detector is trained on 500 ms windows to predict single time points (at 200 Hz sampling rate), our hypothesis has been that there are more than enough examples that are correctly labeled that the model can use to learn from. Indeed, we have seen in past work that this works well, with the speech detector reliably detecting speech events and detecting speech events of varying lengths (reflective of this study's participant's attempts on isolated words of varying length [Metzger et al. 2022, *Nat Comm*; Moses et al. 2021, *NEJM*]). We have added a sentence to the methods section to further clarify our rationale.

We additionally thank the reviewer for raising an important question regarding why a 3.5 second window was used. We used a 3.5 second window given that it was the pace at which our participant could reliably attempt to speak sequential words. At time-windows faster than 3.5 seconds (specifically 2.5 and 3 seconds), the participant was not able to reliably follow the series of cued windows, reporting that he would fall behind as the sentence progressed. 3.5 seconds gave the participant enough time to reliably attempt any word in the dataset. At time-windows longer than 3.5 seconds, the participant felt that there was too long of a break between speech attempts that disrupted his ability to consecutively attempt to speak throughout the phrases. To more quantitatively answer

the reviewer's question, we computed the classification accuracy on isolated speech attempts for windows of different length (Fig. S13). There was no significant difference between 0-3 and 0-3.5s windows; however, a small improvement was seen between 0-3 and 0-4s windows. 0-2.5s windows performed significantly worse than longer windows. Likely, this reflects the fact that the participant does not always start at the cue, sometimes taking longer to start a speech attempt. Having a longer sized window like 0-3.5s allowed the model to consume relevant neural features to the speech attempt, even when behavior was not perfectly aligned to the go-cue. We have further clarified these points in the manuscript. We did not try different sized windows for different words, as the system operates on a fixed cue basis where the window size is constant. The participant found this format to be comfortable and amenable to consecutive speech attempts. However, using variable length windows and an approach that is agnostic to window length, specifically in the context of decoding bilingual phonemes, is an important future direction, and we have added this to the discussion.

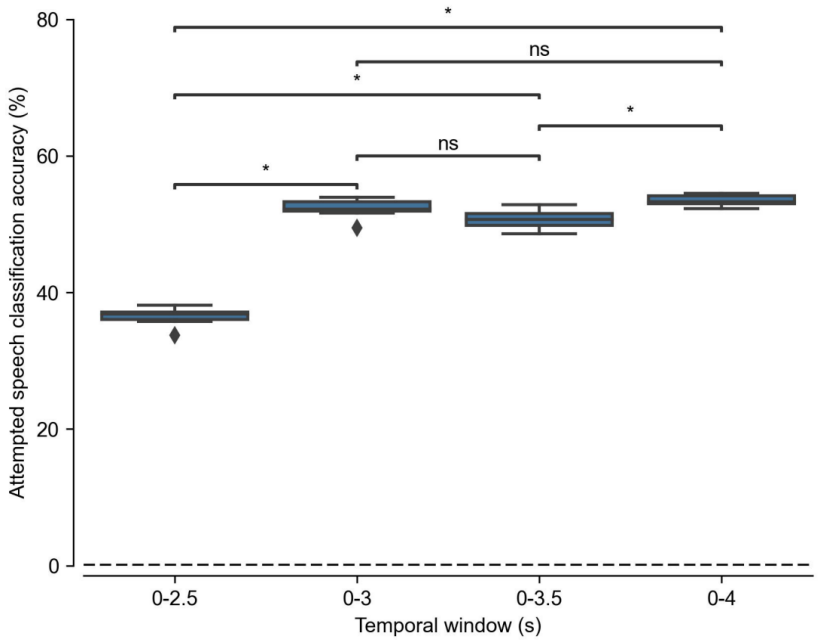


**Figure S13. Performance of attempted speech model using windows of various input lengths.** We evaluated classification performance over the full 104 bilingual words, using models trained with varying length temporal windows of neural features. Classification training and evaluation occurred identically to described in the methodology and each distribution is over 10-fold cross validation (* $P < 0.01$; two-sided Mann-Whitney U-test with 6-way Bonferroni correction for multiple comparisons). Dashed line indicates chance performance (0.96%).

"Time points between the presentation of the target word on the screen and the go-cue were labeled as 'speech preparation' and time points between the go-cue and 2 seconds after the go-cue were labeled as 'speech'. This window was chosen based on the duration of average neural responses during speech attempts. Time points from 2

seconds after the go-cue to the end of the trial (when the screen cleared) were discarded from training, due to the ambiguity of when a speech attempt truly ended and the participant was at rest."

3.5 seconds was chosen for the window length, as it was the speed at which the participant could reliably attempt sequential words using the cued decoding system

We also provide the CV accuracies for using different window sizes during classification of the full 104 bilingual-word vocabulary (Fig. S13).

Here, we demonstrate that, through transfer learning, training time for multilingual participants to use multiple languages may be strongly reduced. Further, our sub-lexical decoders achieved robust performance on syllables in a new language with no additional language-specific training data. These findings suggest that future multilingual speech BCIs may target shared sub-lexical units, for example phonemes 8,14,15,66, to enable rapid generalizability between languages and as an alternate approach to fixed-window word decoding.

6) **Have the authors compared neural activity periods where the subject silently read or listened to the words to periods of attempted speech? It would be important to fully understand whether the attempted speech shared anything with reading or listening to that speech in order to ascertain that the neural representations do indeed reflect motor production of speech**

This is a great control suggestion from the reviewer, and we agree that it would add value to the manuscript. To this end, we collected additional data while our participant attempted to speak, passively listened to, and silently read 10 bilingual words (around 250 trials of each paradigm, with 25 repeats of each word target within each paradigm). We then trained a model on the attempted speech neural features, using the same procedure previously described in the manuscript. To test whether attempted speech shared characteristics with passive listening or silent reading, we evaluated the attempted speech model on neural features from each paradigm. We found that above-chance performance could not be achieved when evaluating on listening or silent reading, in contrast to evaluating on attempted speech (Fig. S4). This provides evidence that attempted speech signals are specifically related to the motor production of speech, at least in the context of this work with the cortical coverage of our implant. We have added text to the results section of the manuscript documenting this result.

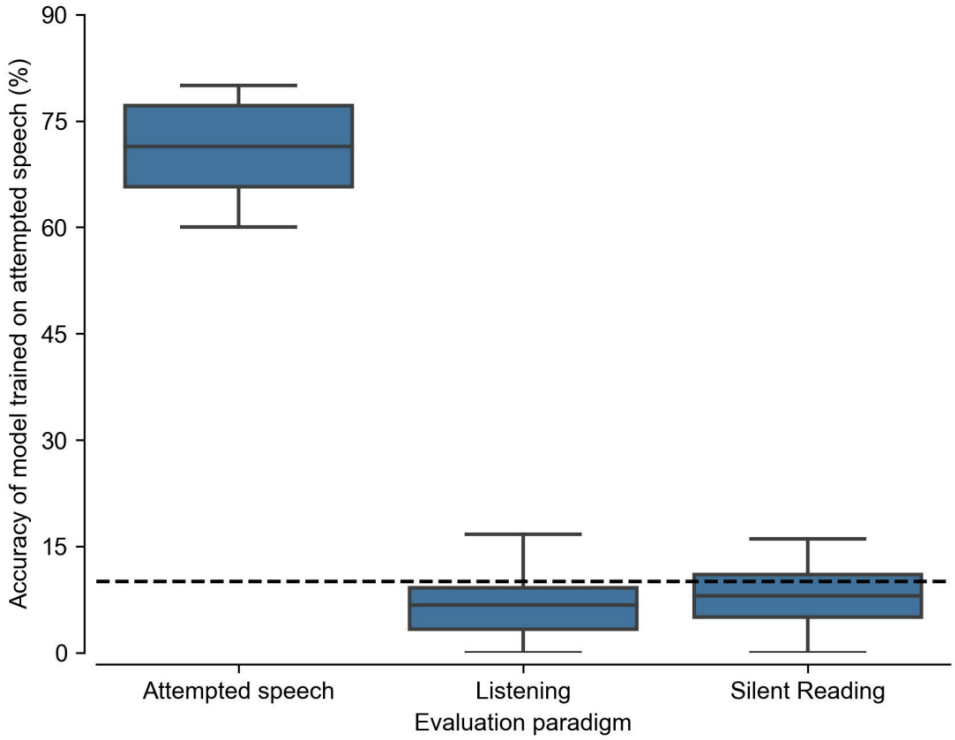


**Figure S4. Performance of attempted speech model on silent reading and listening.** For a subset of 10 bilingual words, we collected neural features during attempted speech, passive listening, and silent reading (roughly 250 trials in each paradigm). A model was trained on attempted speech data, using the same procedure throughout the manuscript, and evaluated on neural features from held-out attempted speech, passive listening, and silent reading trials. Performance was not significantly different from chance when evaluating the attempted speech model on listening or silent reading, in contrast to evaluation on attempted speech. This provides evidence that attempted speech neural features are specific to motor production of speech and not reflecting a process that strongly underlies listening or silent reading. Results are from 10-fold cross validation within each paradigm. Dashed line indicates chance performance (10%).

In line with this result, we verified that neural features were specific to attempted speech and not listening or reading (Fig. S4).

**7) How error in the word error rate (WER) metric was split between neural only versus LLM only, in case of isolated versus phrase contexts?**

We appreciate this question and have added additional clarification to the manuscript. In the case of decoding phrases (Fig. 1), we computed the word error rate (WER) under three conditions: chance, neural-only, and online. In the online condition, the WER was calculated after applying language modeling to the outputs of the neural classifier over the sequential 3.5s windows. This yields a decoded sentence, for example “si yo quiero agua”, which can be compared to a ground truth sentence, for example “si yo puedo

hablar". The WER is just defined then as the edit distance between the ground-truth and decoded sentences divided by the number of words in the ground-truth sentences. The reason that this is different from classification accuracy in an isolated context is that the WER is also sensitive to the number of words decoded. If the decoded sequence has more words than the ground truth, that will be factored into the edit distance. We have added detail regarding this process to the methods section to add additional clarity for the reader. Chance was computed the same way as online except the neural features were randomly shuffled before decoding.

In the case of neural-only, a language model was not applied to the outputs of the neural classifier over the sequential 3.5s windows. Without a language model applied, the verbs remain unconjugated. The most likely classified word from each window is then concatenated together to form the "neurally decoded" sentence (e.g. "yo querer hablar"). This can then be compared in the same way to the ground truth sentence (e.g. "yo poder hablar"), which in this case we also keep unconjugated, to compute a WER using neural data alone. For trials where the length of the decoded and ground truth sentences do not match, we only consider up to the minimum sentence length (between decoded and ground truth) to compute the word error rate. Thus, in the case of neural-only decoding the word error rate is the inverse of classification accuracy in isolated contexts and let us probe the neural-decoding performance only of the system. However, we presented these results as a word error rate to keep consistency with the chance and online metrics on the same plot.

For isolated word classification (not in the context of sentence decoding; Figs 2,3,4) we chose to report the classification accuracy (percent correct) rather than an error rate. In the case of isolated contexts, the classification accuracy is the inverse of the word error rate (since one word is always decoded). We chose to report error rates for the phrase decoding task since that is the standard in automatic speech recognition and in BCI studies that decode neural activity into sentences [Willett et al. 2021, *Nature*; Moses et al. 2021, *NEJM*; Metzger et al. 2022, *Nat Comm;* Willett et al. 2023, *Nature*; Metzger et al. 2023, *Nature*, 2023]. Similarly, we chose to report the inverse, classification accuracy, for isolated tasks since that has been the standard in past studies that decode neural activity into isolated targets [Moses et al. 2021, *NEJM*; Metzger et al. 2022, *Nat Comm;* Wandelt et al. 2022, *Neuron*; Berezutskaya et al. 2023, *J Neural Eng.;* Soroush et al. 2023, *Neuroimage*]

Each block contained 8 phrases, 4 English and 4 Spanish. Similar to prior studies, we choose to report WER over blocks given that short phrases may become overly influential 42,43,75. To compute the WER in the neural-only condition, we left verbs unconjugated both in the decoded and ground truth sentences and considered up to the point where the ground truth and decoded sentences had the same number of words. Thus, it is the inverse of classification accuracy and lets us probe the neural-classification performance of the system during the sentence-decoding paradigm.

**8) Fig2a shows significant differences between English and Spanish word classification rates. Can authors comment on why this is the case.**

We thank the reviewer for this question and agree that additional insight into the difference between English and Spanish classification rates would be helpful to readers. To this end, we computed the mean pairwise mel-cepstral distortion (MCD, as in Fig. 2g) for each word, separately within the English and Spanish vocabularies. We found that, on average, English words have a lower pairwise MCD to other English words than the pairwise MCD for Spanish words to other Spanish words (Fig. S6; P < 0.0001, two-sided Mann-Whitney U-test). This indicates that the English words are more acoustically similar to each other than the Spanish words. Given that the results of Fig. 2g demonstrates that acoustic, and by virtue articulatory, similarity is the biggest driver of confusability, it follows that the English vocabulary should have lower classification rates than Spanish, all other things equal. We have added a supplemental figure documenting this finding and corresponding text to the manuscript to help contextualize any differences between the English and Spanish word classification rates.

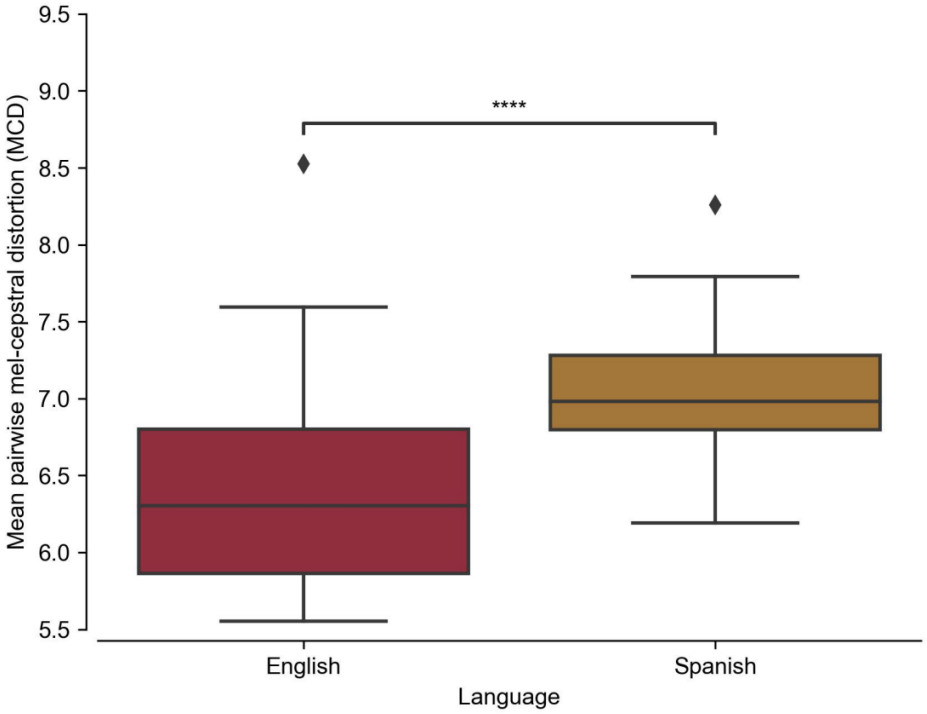


**Figure S6. Acoustic similarity of words within the English and Spanish bilingual words.** For each word in the English vocabulary we calculated the mean pairwise mel-cepstral distortion (MCD) to all other English words. We repeated the same procedure for Spanish. Distributions across words (51 English, 50 Spanish, shared words were excluded) are shown. English words have a significantly lower mean pairwise MCD (**** $P < 0.0001$, two-sided Mann-Whitney U-test). This indicates that English words, on average, are more acoustically confusable with other English words than Spanish words are with other Spanish words.

We achieved median CV classification accuracies of 58.1% (99% CI: [56.9, 59.3]%) overall, 62.9% (99% CI: [61.3, 64.9]%) for Spanish words, and 52.9% (99% CI: [51.6, 55.6]%) for English words (Fig. 2a). We also computed median CV classification accuracy over the full 104-word vocabulary (no masking), which was 47.2% (99% CI: [45.8, 48.2]%; Supplementary figure S1). Differences in English and Spanish classification accuracy may stem from characteristics of the specific stimuli used (Fig. S6)

9) **Fig2: Clarify the extent of decoder training between sessions separated by 30 days breaks**

We agree with the review that this point should be more salient and clarified in Fig. 2. No decoder training occurred between the sessions separated by the 30 day break. The decoder weights for this analysis were frozen at the dotted line in Fig. 2b with no subsequent training. We have clarified this in the Fig. 2b caption.

b, Stable classification of words in English, Spanish, and across both languages for 48 days without retraining or recalibration of the system. Classifiers were trained from data collected over a 50-day period and then the weights were frozen (black dashed line). Results were achieved over 3 years after ECoG-device implantation. The break in the axis indicates a 30-day break in recording sessions with the participant. No retraining occurred between sessions separated by the 30 day break. c, Classification performance before (n=5 days) and after (n=5 days) a 30-day break in recording without retraining.

10) **AAC speed was only based on 8 sentences (4 English and 4 Spanish). This estimate of 6.59 words per minute might be largely biased. Can authors comment on how this could affect significance in Fig1?**

We appreciate this comment from the reviewer as a way to strengthen any comparisons made to AAC speed. 8 sentences is limited and indeed may have been a biased comparison, as many of these 8 sentences had shorter length words than representative of the entire test set. For this reason we re-performed this analysis with 30 sentences: 15 English and 15 Spanish. We also ensured that these sentences had similar word lengths (measured in letters) as the entire online-evaluation set (Fig. S14). After performing this analysis, the new mean estimate with AAC is 3.97 words per minute. We have updated the manuscript to reflect these changes.

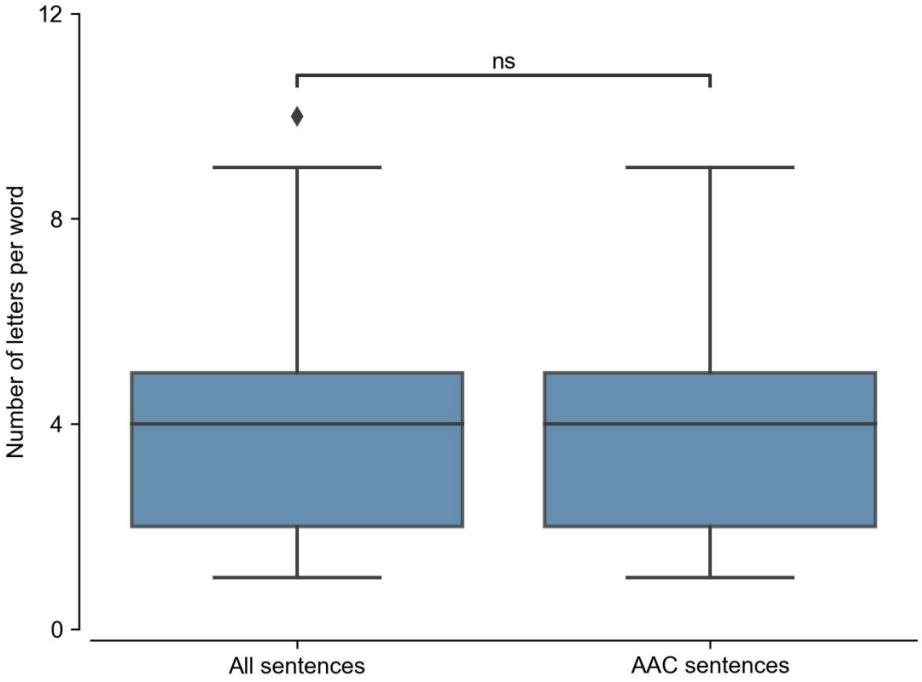


**Figure S14. Comparison of all online-evaluation sentences and AAC evaluation subset.** For each word in each of the online-evaluation sentences, we computed the number of letters. We did the same for the 30 random sentences drawn for the AAC evaluation. Distributions for each are shown with no significant difference (NS $P > 0.01$, two-sided Mann-Whitney U-test)

We quantified the participant's communication speed using his common Augmentative and Alternative Communication (AAC) strategy. In brief, the participant uses residual head movements to control a laser pointer attached to his glasses to spell out words by directing the laser to letters on a board. We recorded the participant spelling out 30 sentences, 15 English and 15 Spanish, that were randomly chosen from the bilingual-test phrases. We then calculated the average words per minute over these 30 sentences. This is the value displayed on Fig. 1d (3.97 words per minute). We also confirmed that the randomly drawn sentences had the same letters/word distribution as the full bilingual-test-phrases (Fig. S14).

**11) It is stated the HFG 70-150 and LFS 0.3-100 Hz were acquired and used as features to classifiers.Given the band overlap, how did this redundancy in the 70-100 Hz band affect classifications? Is there a reason for extending the LFS above 70 Hz?**

We thank the reviewer for this question and agree that more clarification in the manuscript would be helpful. There is no specific reason for extending LFS above 70 Hz; however, we capped LFS at 100 Hz, given that it was derived from the 200 Hz downsampled signal. We partially used these precise cutoffs given their success in

capturing neural features for classification in our prior work [Metzger et al. 2022, *Nat Comm*]. An additional reason that they capture unique information content is that the HGA between 70-150 Hz is in the form of analytic amplitude whereas the LFS 0.3-100 Hz is not. Thus, it is unlikely that in the overlapping band 70-100 Hz, the information is represented identically. We have added additional details to the methods section for how these features were derived.

However, a more important reason for their use is their ability to capture unique features and engage a wide portion of the grid. The electrode contributions for classification are shown for both HGA and LFS in Fig. 2d,e. Looking at the distribution of contributions on the cortex, it can be seen that LFS and HGA engage complementary electrodes. However, to better assess this, we have performed an additional analysis, plotting electrode contributions for LFS against HGA (Fig. S8). Here, we find minimal overlap between the top 10th percentile of electrode contributions for LFS and HGA. This provides further support that these frequency bands capture unique, informative neural features for classifiers. We have added text to the corresponding results section highlighting this result and expanding on the definition of the frequency bands.

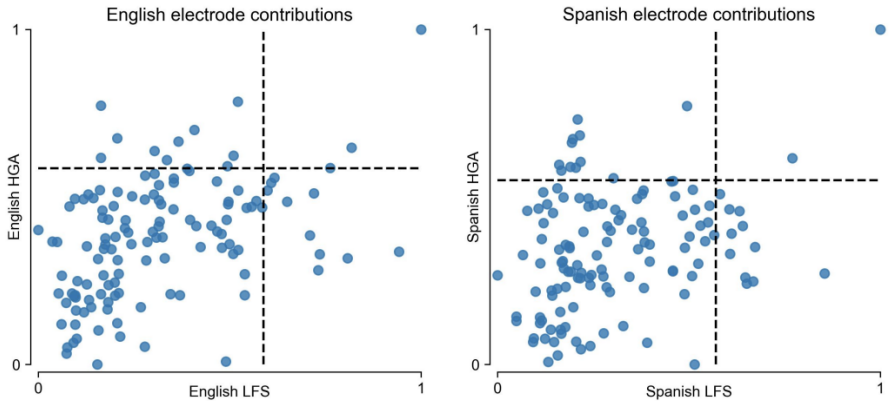


**Figure S8. Distinct contributions of HGA and LFS to classifier performance.** Shown are plots of electrode contributions for HGA against LFS, separately for English (left) and Spanish (right) trained models (as in Fig. 2d,e). The dotted lines indicate the 90th percentile of HGA and LFS contributions. The majority of electrodes only fall above the 90th percentile for one of HGA or LFS.

To extract these features, we used digital finite impulse response (FIR) filters to compute the analytic amplitude of the signals in the high-gamma frequency band (70–150 Hz; HGA) and an anti-aliased version of the signals (with a cutoff frequency at 100 Hz; LFS). We concatenated the time-aligned HGA and LFS into a single temporal stream, sampled at 200 Hz. The HGA is then derived from the analytic amplitude whereas LFS is not. We next, z-scored the HGA and LFS for each channel using a 30-s sliding window approach.

Indeed, non-parametric correlation between electrode contributions for English and Spanish, for both HGA ($P < 0.0001$, $\rho=0.85$, non-parametric correlation permutation test; Fig. 2d) and LFS ($P < 0.0001$, $\rho=0.92$, non-parametric correlation permutation test; Fig. 2e), showed a strong positive relationship. In contrast, for both English and Spanish, very few electrodes contributed strongly with both LFS and HGA, indicating complementary information in the two bands (Fig. S8)

**12) Why did they have to train the classifiers on silent isolated word attempts before they started to retrain on the attempted grunts/moans? How was the detector threshold developed and fine tuned in such case?**

We appreciate the opportunity to clarify this aspect of the classifier training. Prior to beginning collection of the overtly attempted dataset for this manuscript, we had collected a limited silently attempted task with the bilingual-words. In our prior work [Metzger et al. 2022, *Nat Comm*], we demonstrated that transfer learning between silently attempted speech and overtly attempted (grunts/moans) speech attempts improved classifier performance. We thus attempted to expedite training of the overtly attempted classifiers by using the prior silently attempted dataset for pre-training. Ultimately, this increased performance by a large amount early during data collection, when, for example, 10% of the total overt data was collected (Median 21.00% improvement, $P<0.0005$, two-sided Mann-Whitney U-test), and by a small amount (Median 3.12% improvement, $P<0.0005$, two-sided Mann-Whitney U-test) when all total overt data was used for training. Given these improvements to performance, we chose to leverage the prior silent isolated word attempts dataset to help train our final overt models. We have added a supplemental figure (Fig. S12) comparing performance with/without the mimed-pretraining as well as more detail to the methods section to improve clarity for the reader. The detector was only trained on overtly attempted speech attempts and did not use any silent speech attempts in the training data. We apologize for a lack of clarity here. We have updated the manuscript text to better reflect these points and to reference the new supplemental figure.

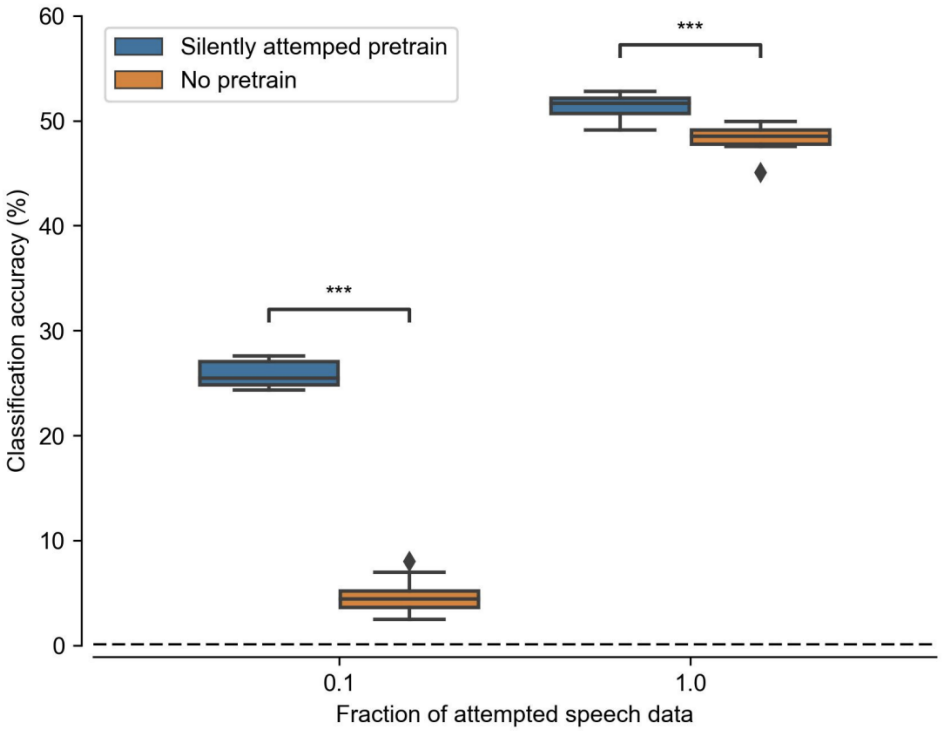


**Figure S12. Effect of pre-training with silently attempted speech data.** Shown is performance of classification models, over the full 104 bilingual words, at two fractions of total attempted speech data: 0.1 and 1. At both fractions, pre-training with the limited silently attempted speech data helps improve performance, albeit marginally with full use of the attempted speech data (Median 21.00% and 3.12 % improvement at 0.1 and 1, respectively, *** $P < 0.0005$, two-sided Mann-Whitney U-test). Dashed line indicates chance performance (0.96%).

We trained an artificial neural network (ANN) with the objective of classifying the neural features from a speech attempt as one of the 104 words in the bilingual-words set. During model training, we used stochastic gradient descent and the Adam optimizer 77 to find a set of parameters that minimized the cross entropy loss between target and predicted outputs across the 104 classes. Weights were initialized with parameters learned on the prior collection of mimed bilingual-words, given a small, final improvement on overt with transfer learning from mimed (Fig. S12) .

In this work, we train the model on overt isolated-target bilingual-words data and rest blocks (where the participant rested silently for 1 minute). For each isolated-target bilingual-words block, we labeled each time point as ‘rest’, ‘speech preparation’, or ‘speech.’ Time points between the presentation of the target word on the screen and the go-cue were labeled as ‘speech preparation’ and time points between the go-cue and 2 seconds after the go-cue were labeled as ‘speech’.

## Reviewer 2

1) **The study by Silva et al. from the Chang lab is an extraordinarily clear and clean demonstration, both of a brain function and of a technical small masterpiece: The real-time, soft- and hardware-aided decoding of language and language content from sensori-motor, articulatory (i.e., neural) representations in the human brain, which allows the authors to promote this as a demonstrator of a bilingual speech neuroprothesis. For once, this is hardly an exaggeration.**

   **The manuscript is technically advanced but for most uses established neural encoding/decoding routines long established by the lab and other ECoG/invasive-ephys researchers, combined with the power of large language networks (here, GPT-2 and word2vec if I am not mistaken). The strength of this manuscript and its novel aspect thus clearly lies in now approaching the question of bilingual speech representation and the feasibility of bilingual neuroprotheses.**

   **I consider this m/s as presented a fortunate instance of a two-way street between neuroscience and neuroengineering. A long-standing question on bilingual representations gets tackled and in part answered by the data presented (see eg. Fig. 3), but doing so results from a more engineering-geared question asking why the bilingual decoder works in the first place.**

   **I have been pondering the major or minor concerns one could or should additionally bring forward against the manuscript in its current form. However, due to the fact that this is by all intent a proof-of-concept report and not an inference-on-population study, the criteria are clearly different. For a single, highly-specialised patient this is by all means an impressive, and methodologically clear report with neuroscientific and neuroengineering value.**
   **While different labs might have taken different analytical choices, the report as is is transparent and sound enough to clearly allow the reader to judge for themselves. Minor quibbles on why the outdated/underpowered Wilcoxon/Mann -Whitney (instead of permutation tests) keeps being should thus not be a point of contention here.**

   We thank the reviewer for their positive and encouraging summary of our work and the great suggestion below to include an additional discussion paragraph highlighting implications for non-invasive neuroscience and generalization to other patient groups.

2) **I would recommend that the Discussion section is amended with a paragraph on potential roads to application in other patients or its applicability to non-invasive neuroscience (citing e.g. Correira et al., 2014, https://www.jneurosci.org/content/34/1/332).**

We thank the reviewer for this suggestion and agree that adding a paragraph to these effects would improve the manuscript. We have added a paragraph to the discussion section touching on both roads to application in other patients, including what language-related factors should be tested in future studies and applications for non-invasive BCIs. The discussion explicitly discusses the suggested work (Correira et al.) as an extension of the described framework in our study.

Although this is a single-participant study, the strength of shared cortical-articulatory representations is encouraging for generalizability to other patients. Notably, as this participant learned L2 (English) later in life, this may be even more encouraging for those who learn L2 early in life, which has been shown to lead to stronger shared representations 30,31,33,35. Future work, however, should directly investigate whether this effect varies with L2 proficiency, age-of-acquisition, and articulatory similarity to L1. While this study leveraged invasive electrophysiology, the results also hold relevance for non-invasive BCIs. A potential advantage of non-invasive BCIs is the ability to sample a larger distribution of cortex involved in functions beyond articulation. Non-invasive BCIs may apply a similar framework, as applied here with articulatory features, for other shared language features, such as semantics 67,68, that could enable rapid generalizability of a BCI across languages. These studies may also find language-specific signals, which we did not observe in speech-motor cortex, that may exist in higher-order cortical regions 28,69,70. A complementary direction for invasive BCIs is to study whether language-specific signals exist at the level of single neurons.

## Reviewer 3

1) **Silva et al. present a study that extends a prior proof of concept brain to text system (Moses et al., NEJM, 2021) from English only to a bilingual (English and Spanish) system. The core demonstration and additional analyses will be valuable to a burgeoning field and provides forward vision towards the broader translation and use of this technology. Their core demonstration is that the same prosthesis can be utilized for the L1 (Spanish) and L2 (English) languages of their participant and that transfer learning approach (commonly utilized by the broader machine learning community) can be successfully employed to more rapidly train and adapt the neural prosthesis to a second language. Their findings, consistent with our expectations from prior literature, suggest that the overlap in neural activity present in both languages that enables their technical demonstrations is the result of a shared articulatory representation for both languages.**

   **Although there is only one participant in this study, given the challenging nature of this work and the advance in knowledge even an N of 1 provides in this field, it**

**should be strongly considered for publication. N of 1 publications in this field are increasingly common, for example, (Moses et al., NEJM, 2021), (Willet et al., Nature, 2021), (Willet et al., Nature, 2023), (Metzger et al., Nature, 2023).**

We thank the reviewer for their summary of our work and for recognizing the value of single-participant studies in the context of brain-computer interface clinical trials.

2) ***Throughout the manuscript the term "real time" is used. For a broad bioengineering audience, the use of this term could lead to confusion. "Real time" for many engineered systems implies microseconds or milliseconds of latency. It can also be interpreted as a system with output delay (latency) that is low or negligible for the given application. In the case of a speech and language system there are likely a few key numbers that could be considered — for example if this was a speech synthesis system a delay of 10s of milliseconds could be argued to be real-time, as such a delay introduced for able bodied speakers does not result in appreciable speech deficits. Since this is a text based output system, I think it would be reasonable to look at the median turn taking time between two healthy individuals, which is on the order of 100s of milliseconds. However, if I understand your system architecture correctly (and from watching the supplementary videos), the latency is multiple seconds. Thus, I think real time is not an appropriate description and could lead to misunderstanding by many readers. It might be better to describe this system as operating "online." If you do use "real time" it should always be qualified with the effective latency between speech intent and output to screen.***

The reviewer raises an excellent point. We agree that the term "online" is a better description of the processing occurring in our text-based speech neuroprosthesis. We have changed "real-time" to "online" throughout the manuscript.

3) ***Related to the previous concern, the schematic of the decoder in Figure 1a may mislead some readers. The use of 3.5 seconds of data along with a bi-direction RNN could be better highlighted. The current figure could be misinterpreted to be a uni-directional RNN with shorter input segments. I'd suggest augmenting Figure 1a to more clearly show the flow of data / relative timing. I'd also suggest editing the Supplementary Methods section S1 and adding a Supplementary Figure that more fully describes the algorithms/architecture and temporal flow of data.***

These are helpful suggestions from the reviewer that parallel the prior comment. We have re-made Fig. 1a to emphasize that the RNN is bidirectional and takes the full context of 3.5s windows. We also added labels to each 3.5s window and additional clarification that we hope makes it more clear that the process shown on the right of Fig. 1a occurs after each 3.5 second window. In addition, we have also created a figure (Fig.

S1) that describes how, for a given trial, information flows from the neural features through the decoding processes and finally to the participant monitor. We hope that this better helps the reader contextualize the supplementary videos and the overall flow of information through the system. To allow the reader to better understand the architecture and steps in the classification models, we have also created a graphical depiction of the word-classification process (Fig. S2). To complement these figures, we have added additional clarity and edits to the supplementary Methods section S1, as suggested by the reviewer. Here, we specifically added a section that outlines, step by step, the information processing and flow for a given trial that references Fig. S1.

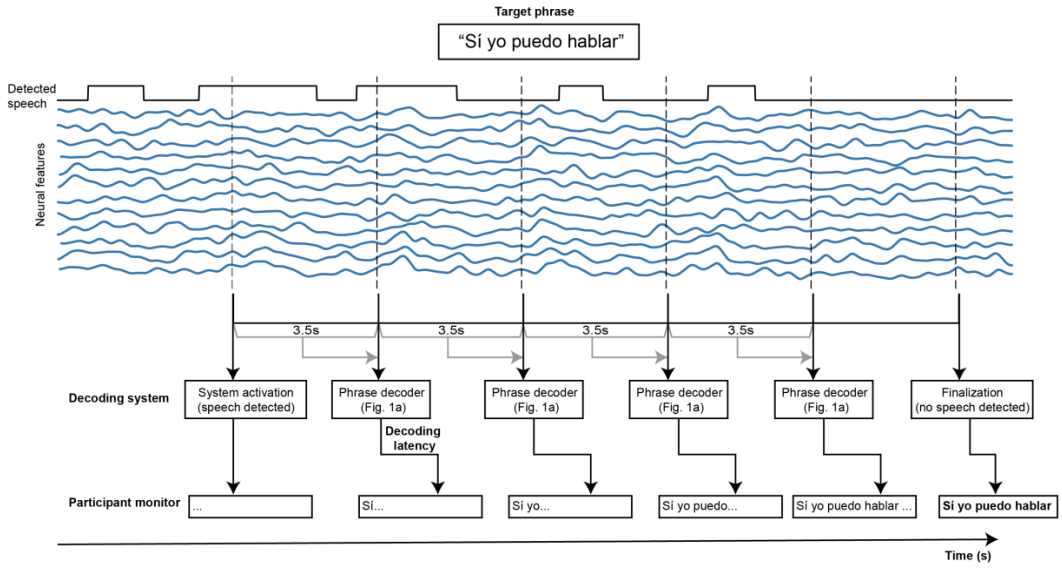


**Figure S1. Timing and information flow through the bilingual-sentence decoding system.** Shown is a more detailed schematic overview of the bilingual-sentence decoding system to complement Fig. 1a. Three levels of information are depicted: the neural features, the decoding system, and the output to the participant monitor. To start, the participant makes a speech attempt. This is detected by the system and cues activation of an ongoing decoding process. Following activation, a series of 3.5s windows are cued to the participant. At the end of each window, after the full 3.5s have passed, the neural features from that window are passed to the decoding process illustrated in Fig. 1a. Following a latency to conduct the decoding, the most likely beam from the process in Fig. 1a is displayed on the participant monitor. This process continues to occur for sequential 3.5s windows until a window with no detected speech occurs. After such a window, the decoding is finalized and terminated. The system then listens for another speech attempt to activate and repeat the process.

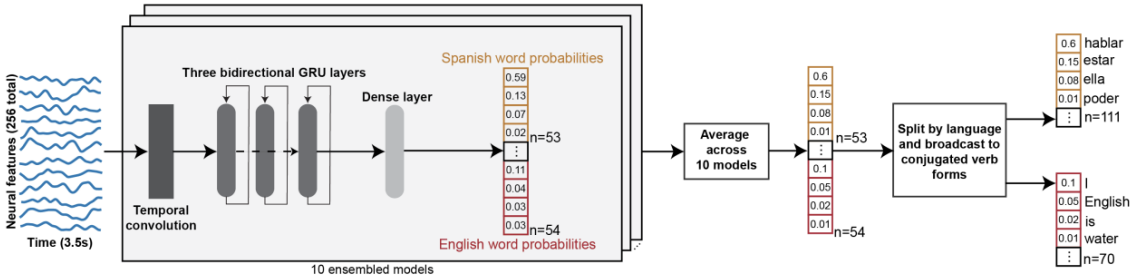


**Figure S2. Graphical depiction of bilingual-word classification.** Shown is a schematic of the bilingual-word classification process. Neural features (256 total; 128 HGA and 128 LFS time series over 3.5s) are classified as a word in the bilingual vocabulary. Neural features are first processed by a temporal convolution. Next, the features are passed through three bidirectional GRU layers. The latent state from these layers is then read out by a dense, linear layer that emits probabilities over the 104 words in the bilingual vocabulary. This process is performed by 10 distinct models, each with a different weight initialization and trained on different folds of the data. The probabilities generated across these 10 models are averaged to create one probability vector across the bilingual vocabulary. This vector is finally split by language and the probability for a given word is broadcast to all conjugated forms of the word before being combined with the language model, as shown in Fig. 1a.

The phrase-decoding system was designed to decode English and Spanish phrases from cortical activity without the user having to manually select a language. The system continuously acquired neural features, high-gamma activity (HGA; 70-150 Hz) and low-frequency signals (LFS; 0.3-100 Hz), from the local field potential (LFP) of each electrode on the participant's array. The temporal flow of information through the system is as follows (Fig. S1). The participant first makes a speech attempt which is detected by the system, through the speech detection model, and cues activation of an ongoing decoding process. Following the system activating, sequential 3.5s windows are shown to the participant. At the end of each window, after the full 3.5s have passed, the corresponding neural features from that window are passed to the decoding process illustrated in Fig. 1a and Fig. S2. There is then a latency to conduct the beam search and decoding, after which the most likely beam from the process is displayed to the participant monitor. This will continue to occur until a 3.5 window passes were there is no detected speech event. After such a window, the decoding is finalized and terminated. The system then resets and waits for another speech attempt.

4) ***I think the definition of "stability" needs to be better qualified. The comparison to intracortical techniques and results are a bit confusing and misleading. In this case, the authors are highlighting system performance without daily recalibration. However, for the field it is important to separate the stability of neural features (i.e., stable relationship between neural modulation and intended language output) vs. stability in the information content in the neural features (i.e., relationship***

***between neural modulation and intended language output is not stable, but overall information content is stable). For this work, it is interesting to see that performance does not trend downwards when the same decoder is used from day to day, but the word "stable" is misleading, as the graphs show quite a bit of variability. There is about a 20% difference between the top performing and lowest performing day — I don't think it's reasonable to consider decoding performance of 50% - 70% to be "stable." If we compare this variability to the recent Willet et al, Nature, 2023 study, one could argue that their intracortical result is far more stable — Fig 2b in that paper shows a much greater stability during online control sessions (~5% difference between lowest and highest performing days) for both a similar complexity task (50 word vocabulary) and higher complexity task (125,000 word vocabulary). Of course, an important caveat in that study is that they are retraining the decoder everyday, whereas in this study the decoder was fixed. A valuable course of action would be to compare performance when retraining each day to performance when holding the model fixed. Since all of the characterizations Figure 2 of this manuscript are offline, these analyses should be straightforward to conduct. Then the results should be presented with proper qualification of the word "stable" under these two very important contexts.***

The reviewer has raised multiple helpful points to heighten the clarity and improve the interpretation of the results presented in Fig 2b-c. First, we agree that a direct comparison with intracortical techniques is a bit out of place in this context and not critical to the core results or interpretation. For this reason we have removed such comparisons from the manuscript.

We also agree that our presentation of the results as "stable" may not be immediately clear and deserves further workup and description under the two definitions proposed by the reviewer. Please correct us if we are misunderstanding, but by the first point we feel the reviewer is asking if the neural modulation to attempted speech is consistent for each utterance over time (i.e. attempting the word "hello" would evoke the same pattern A over time whereas attempting the word "goodbye" would evoke the same pattern B over time). By the second point, we interpret the reviewer's comment to mean that the neural signals consistently contain information content to discriminate speech utterances but the evoked patterns for utterances may change over time. Our main goal with this analysis in Figure 2 is to claim the first point: that stable neural modulation exists in the neural features, over the time course of months, allowing models to maintain performance without re-calibration. We have edited the wording of the paragraph to better highlight that this is the main claim. We feel that this claim is supported given that the model is able to retain the same performance over months (Fig. 2b) without re-calibrating the specific neural modulation patterns that the model has learned to be associated with each word. However, we acknowledge that this may be accompanied by some day-to-day variance in the signals. We conducted the analysis suggested by the reviewer to compare performance when retraining each day to performance when holding the model fixed (Fig. S7). We note that moderately improved performance is

possible when retraining each day (Fig. S7). This may reflect the benefits of adding additional data to the models. However, there is still a decent amount of fluctuation in model performance day-to-day, even with retraining, indicating a baseline level of variance in the signals. Thus, while we believe that point 1 is supported, it is accompanied by some variance in the neural signals. We have clarified in the results section that our goal is not to claim minimal performance variance day-to-day but instead that a model can pick up on the same patterns in the neural features over months.

We have re-written the corresponding results section to incorporate these results and clarify the intention of our analysis and what is meant by stable model performance. Here, we most strongly emphasize our main claim that models may retain the same performance over months without the need for re-training, due to, as the reviewer notes, stable neural modulation patterns to the different words over time. We also highlight that, given performance is achieved over 3.5 years after device implantation and comparable to our first study with the participant [Moses et al. 2021, *NEJM*], there is also longevity in neural information content related to attempted speech (related to the reviewer's second point). Finally, we changed the implant date to correspond to the surgical implantation date of the participant rather than the first day of data recording with the participant, since our claims relate to how long the hardware has been implanted.

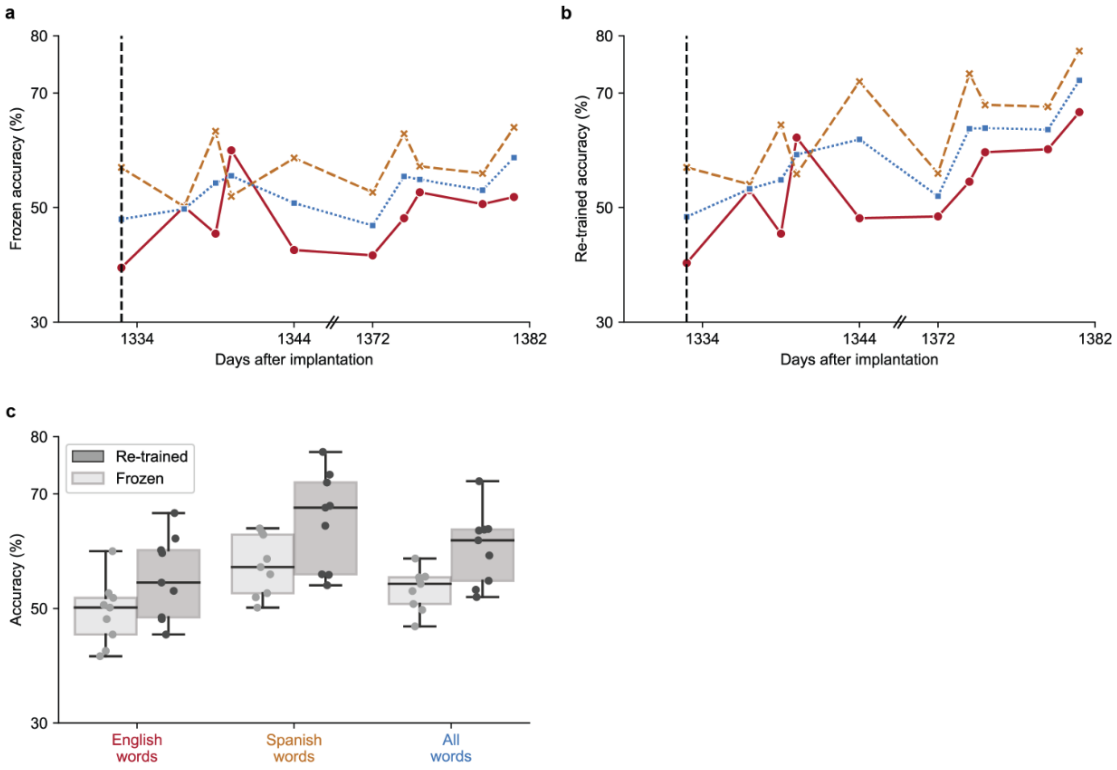


**Figure S7. Effects of re-training models daily during frozen-decoder evaluation.** Shown is a comparison between performance with and without re-calibration. **(a)** Shown is the performance without re-calibration for reference taken from (Fig. 2b). **(b)** Shown is the performance with re-training the classifier with sequential addition of each day's data. **(c)** Shown are distributions of accuracy with and without re-training, demonstrating that small improvements may be found with re-training the decoders with each day's data. Chance is 1.85% for English, 1.89% for Spanish, and 1.87% for all words (masked).

An important challenge in developing clinically viable BCIs is maintaining similar decoding performance day-to-day without requiring the user to dedicate time to frequently recalibrate the system. The relatively large spatial-sampling scale of ECoG offers the potential for consistent day-to-day signal acquisition 45–49. Notably, we found that our models were able to maintain similar classification accuracy, with some day-to-day variance, for 48 days without recalibration (Fig. 2b,c; Fig. S5). This highlights that similar modulation patterns for each word in the neural features are maintained over 48 days. These results were achieved over 3.5 years after ECoG-device implantation with improved classification performance on a slightly larger vocabulary than initial work with this participant 13, demonstrating longevity of speech-information content in the neural signals. Modest improvements in performance were possible with adding more data to the models by re-training each day (Fig. S7).

5) ***Is chance performance calculated with the language model in place? I believe this is the case based upon line 186 in the manuscript. Given that the chance calculation is based upon a shuffle and the core metric of the work is a word error rate, I would like to see a "neural activity only" equivalent for testing chance. This would allow the reader to better disambiguate the role of real-time neural observations vs. integration across time through application of a language model. An understanding of this tradeoff will be a critical design criteria for practical systems and for considering a tradeoff between system latency, vocabulary size, and performance. Why is chance error sometimes greater than 100%?***

We thank the reviewer for this question and appreciate the chance to elaborate and supplement our computation of chance-system performance. The reviewer is correct that the chance performance in Fig. 1b is calculated with the language model in place and with the neural predictions generated from shuffled neural features. We agree that supplementing this analysis with a neural-only chance measure would be useful. To compute neural-only decoding performance of the classifiers, we compute the error rate between the unconjugated ground truth sentences and the unconjugated decoded sentences. For trials where the length of the decoded and ground truth sentences do not match, we only consider up to the minimum sentence length (between decoded and ground truth) to compute the word error rate. Thus, in the case of neural-only decoding the word error rate is the inverse of classification accuracy. We have added detail to the methods section regarding this approach (also discussed in response to Reviewer 1 comment 7). Thus, in order to compute a neural-only chance distribution, we simply shuffled the neural features and passed them through the classifiers to generate chance predictions for each word. We then computed the word error rate the same way as described above. We have added these results as a supplemental figure (Fig. S3) and across all, English, and Spanish phrases, neural-only decoding performance is significantly better than chance ($P < 0.0001$, two-sided Mann-Whitney U-test with 3-way Holm-Bonferroni correction).

The reason that chance performance in Fig. 1a may sometimes be greater than 100% is because it is computed over the full performance of the online system. Thus, in some trials there may be more words decoded than in the ground truth sentence, due to an extra detected window. For example if the ground truth for the trial was "yes I can go" and the decoded sentence was "no please not now or" the word error rate would be 1.25.

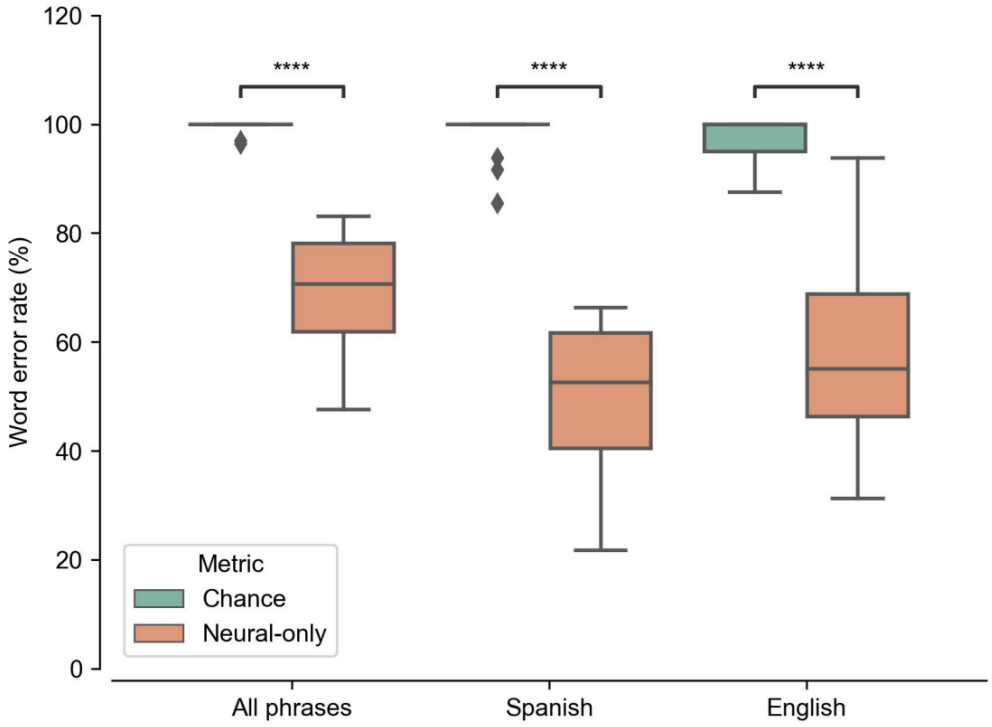


**Figure S3. Neural-only chance sentence-decoding performance.** Shown are neural-only specific chance sentence-decoding distributions, alongside the neural-only decoding performance shown in Fig. 1. Here, we specifically computed a chance distribution with respect to neural-only decoding. We did this by shuffling the neural features and passing them through the classifier. The chance error rate was then computed the same way as for neural-only performance (**** $P < 0.0001$; two-sided Mann-Whitney U-test with 3-way Holm-Bonferroni correction for multiple comparisons).

Each block contained 8 phrases, 4 English and 4 Spanish. Similar to prior studies, we choose to report WER over blocks given that short phrases may become overly influential 42,43,75. To compute the WER in the neural-only condition, we left verbs unconjugated both in the decoded and ground truth sentences and considered up to the point where the ground truth and decoded sentences had the same number of words. Thus, it is the inverse of classification accuracy and lets us probe the neural-classification performance of the system during the sentence-decoding paradigm.

Decoding with neural data alone, without language modeling or beam search, achieved median word error rates of 70.6% (99% CI: [61.9, 78.1]%) overall, 52.5% (99% CI: [40.4, 61.7]%) for Spanish phrases and 55.0% (99% CI: [46.3, 68.8]%) for English phrases (Fig. 1b), indicating that decoding did not only depend on the use of LMs (see Fig. S3 for a comparison to chance neural-only performance).

6) ***The transfer learning result is tantalizing. However, the workup of the result is quite limited and would be greatly strengthened by evaluating the shared aspects of the two languages that facilitate transfer learning — the results shown suggest that pre-training with either language results in equivalent reductions in required training set size. Training set size is a major consideration for the design of these systems — as shown in figure 3, without pretraining 6+ hours of subject specific training data are required to achieve performance that approaches the asymptote. How many hours of data were used for pre-training?***

***To better understand what shared aspects are at play, I would suggest that the authors examine the role of phonemic overlap / articulation feature overlap between the pre-train dataset and the target dataset for transfer. Examination of the impact on transfer can be conducted within language or across languages. Overlap can be controlled by sub-selecting phonemic or articulation features to alter the distributions shown in Supplementary Figure 3. This sub-selection could be conducted by design or randomly and the degree of overlap could be evaluated by measuring the KL divergence between the distributions of phonemic / articulation featured in the two datasets. Such a workup could provide better insight into the bilingual results of these paper and could also be helpful when considering transfer to other languages (and could help add to the wonderful discussion point when considering Mandarin). Furthermore, such a workup could also inform the construction of effective and efficient training sets for single language speech prostheses.***

We thank the reviewer for these excellent suggestions and agree that the workup of the transfer learning result can be strengthened. To this end, we conducted two additional analyses on the two topics suggested by the review: the effect of hours used to pre-train models and the effect of articulatory similarity between the pre-train and fine-tune/test set. In working up the second domain, we chose an approach that would also provide more insight into a question from Reviewer 1 (#4), regarding how well models would generalize to words that were *unseen* during training. We also have corrected an error in the first version of the manuscript where hours of training data was calculated off the entire (104 word dataset) instead of from the sets of 25 words used in Figure 5. The training times are now significantly lower (roughly divided by 4).

To answer the reviewer's first question, we computed the learning curves, under the paradigm of pre-train on English and fine-tune/test on Spanish, using varying amounts of data for the pretrained English models. We have added a supplementary figure (Fig. S11) that plots the learning curves as a function of the amount of data used in the English pretrained models. For the reader's clarity, we also added a second panel which, instead of showing the full learning curves, shows the distribution at select hours of fine-tune/test data for all amounts of English pre-train data. Overall, this analysis supports that more data in the pretrained model improves the rate at which the fine-tune/test model can increase performance; however, this effect plateaus, as there is

minimal difference between using 3.5, 3.9, or 4.3 hours of pretrain data. We have added text to the transfer learning section highlighting this result and referencing the supplemental figure. We have also made clear that for the plot in Fig. 5b, 4.3 hours of data was used to pre-train the models.

To work up the reviewer's second, interesting suggestion, we leveraged a slightly different approach. We chose this approach as we wanted to ask an additional question, noted by reviewer 1: how well can a model perform on words that were unseen during training? To this end, we designed a single set for which to test models. We then chose two additional sets to minimize and maximize the articulatory similarity to the test set (using the same MCD metric as in Fig. 2 for consistency). Interestingly, we found that when pre-training on the set with stronger articulatory similarity, learning curves not only increased more quickly, but we saw high decoding performance on the new words without fine-tuning at all. Thus, the model is able to generalize to articulatorily similar words with no additional training data. We found this result to be a nice extension of the main claims of the paper and, as the reviewer alluded to, potentially informative for monolingual or bilingual utterance set design. We have added these results to Fig 5 as panels c, d, and e. We have also added a paragraph to the corresponding results section discussing these findings. We highlight 1) that pretraining on an articulatory similar dataset leads to faster improvement in learning curves and that 2) if the fine-tune/test set is chosen to match the articulatory characteristics of the pre-train set, it is possible to achieve significantly above chance decoding with no additional training data that is specific to the fine-tune/test set.

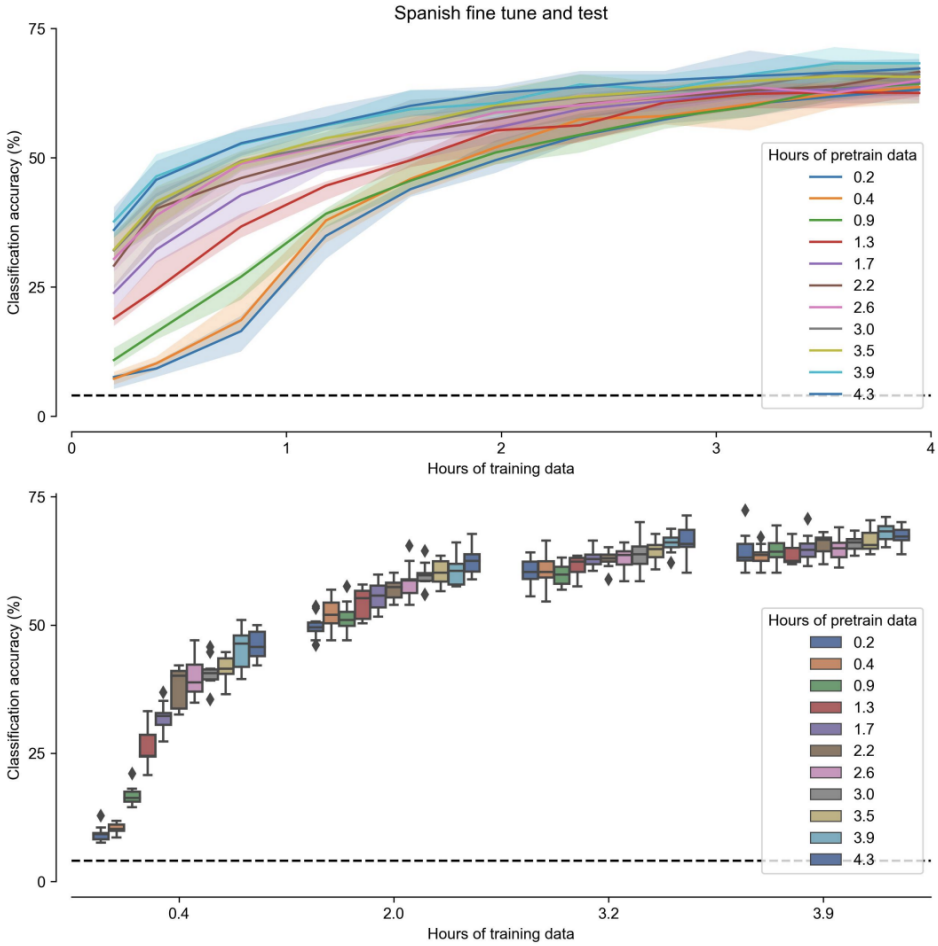


**Figure S11. Effect of amount of pretrain data on transfer learning efficacy. (top)** Shown are learning curves on the Spanish vocabulary (Fig. 4b) with transfer learning from English models trained on varying amounts of data. The amount of hours of pretraining data has an effect on transfer learning efficacy; however, this effect saturates. **(bottom)** Shown is the classification accuracy at select amounts of training data, again as a function of the hours of data used to pretrain the English models. The distribution of accuracy across hours of pretraining data is easier to visualize.

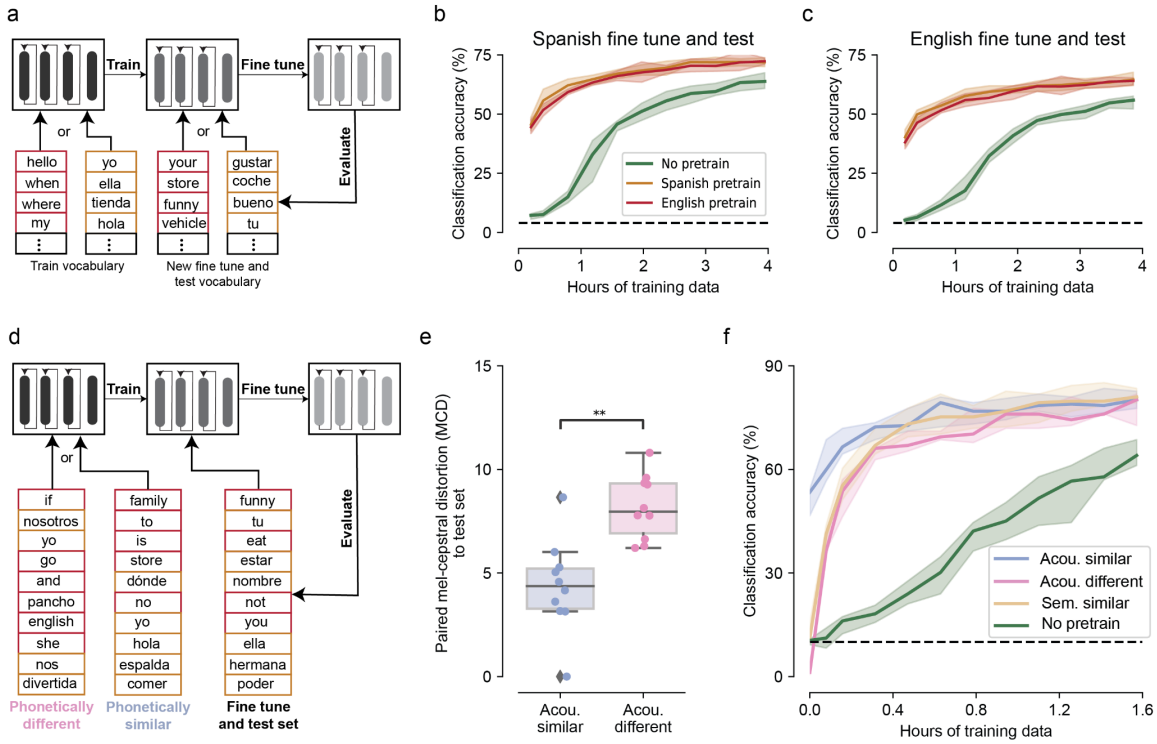


Figure 4. Rapid transfer learning between languages. **a**, Schematic depiction of the paradigm used to evaluate transfer learning between languages. Models were trained on an English or Spanish vocabulary. These models were then fine-tuned and evaluated on a new English or Spanish vocabulary. **b**, Learning curves for fine-tuning and evaluating on a new Spanish vocabulary. **c**, Learning curves for fine-tuning and evaluating on a new English vocabulary. In **b–c**, models were either not pre-trained or pre-trained on a different English or Spanish vocabulary. Chance decoding level is 4%. **d**, Schematic depiction of the paradigm used to evaluate the effect of acoustic similarity between the train and fine-tune set on transfer learning efficacy. **e**, Mel-cepstral distortion (MCD, as in Fig. 2g) between each word in the acoustically (acou.) similar or different train sets with the corresponding word in the fine-tune/test set (** $P < 0.005$; two-sided Wilocxon Signed-rank test). **f**, Learning curves for fine-tuning and evaluating on the “Fine-tune and test set”, defined in d, with transfer learning from the acoustically (acou.) similar and different models. Learning curves are also shown for no pretraining and transfer learning from a model trained on a semantically (sem.) similar set of words. Performance at 0.0 hours of training data reflects the pretrained model’s ability to generalize to the fine-tune and test set with no additional or specific training data. Chance decoding is 10%. In **b–c** and **f** the shaded areas represent 99% confidence intervals.

We pre-trained English and Spanish models on 4.33 hours of data from a vocabulary of 25 words in each respective language (for an analysis of how the amount of pre-training data affects transfer learning see Fig. S11.

To further explore the factors that drove transfer learning and whether models could generalize to an entirely new vocabulary, with no additional training data, we designed a new 10-word fine-tune and test set (Fig. 4D). We next designed three pre-training sets by choosing a word to pair with each of the 10 words in the test set that was either acoustically similar, acoustically (acou.) different, or semantically (sem.) similar (Fig. 4D). We verified that the pairwise MCD between each word in the acoustically similar set and

the corresponding word in the test set was significantly lower than that of the acoustically different set and the test set (Fig. 4E; $P < 0.005$; two-sided Wilocxon Signed-rank test). We then computed learning curves on the new test set, using transfer learning from one of the three pre-training sets or no transfer learning (Fig. 4F). Here we also include an evaluation point at 0 hours of training data, which indicates the ability of the pre-trained model to generalize with no additional data specific to the fine-tune set. Notably, we see that only in the case of an acoustically similar pre-training set is this possible, and learning curves with the acoustically similar set increase most rapidly (Fig. 4F). Overall, this demonstrates that careful design of pre-training and fine-tune/test sets for acoustic, and by virtue articulatory 2,54, similarity may allow generalization of models to bilingual vocabularies with as little as no additional training data.

7) ***The introduction is missing references to similar studies that employ sEEG and intracortical electrodes. These should be included where appropriate in the introduction. Further, the implications for all implanted speech prosthesis should be discussed. Many of the findings in this work likely have implications for recent intracortical electrode based systems, as the system designs have many parallels, but also some unique aspects (e.g., lower effective latency in intracortical demonstrations / orders of magnitude larger effective vocabulary). This is also an opportunity to discuss some of the important subtleties that are critical design tradeoffs for these systems: achieved rate (which may also tie into natural speaking rates of the user and effective vocabulary size / limits to output flexibility caused by language model priors).***

We thank the reviewer for raising these points that would help strengthen the broad appeal of the manuscript. We have made a few changes to these ends. First, we have modified the introduction to specifically highlight sEEG, intracortical electrodes, and ECoG as invasive methods that capture neural activity relevant for speech. Here, we have added 2 new sEEG citations [Tessy et al. 2023, *J Neural. Eng.*; Soroush et al. 2023, *NeuroImage*] and 3 new intracortical citations [Stavisky et al. 2019, *eLife*; Willett et al. 2023, *Nature*; Wandelt et al. 2022, *Neuron*] to complement the existing citations in the manuscript. We also added these citations and our recent work [Metzger et al. 2023, *Nature*], where appropriate, to the following sentence which describes a focus on monolingual decoding

In addition, we have added a new discussion paragraph incorporating the reviewer's points regarding implications and tradeoffs concerning other recording modalities and approaches. We specifically link our sub-word decoding results to a potential for future approaches that explicitly target shared bilingual phonemes to expand both vocabulary sizes and speeds by avoiding the need for sequential cued windows [Willett et al. 2023, *Nature;* Metzger et al. 2023, *Nature*]. We also agree that it is critical to discuss the implications for all speech neuroprosthesis, both invasive and non-invasive. To this end, we have added another paragraph discussing how the results may be complemented by

non-invasive and future invasive, specifically single-unit, studies. Here, we mention the potential to find language-specific signals in either single-neurons or in higher-order cortical regions. Like the reviewer states, this would mitigate the need for the bilingual neuroprosthesis to accumulate linguistic context from the language model before decoding the correct language. We hope that these changes have addressed the reviewer's points, although we are happy to edit further if desired.

Specifically, intracortical electrodes, stereo-electroencephalography (sEEG), and electrocorticography (ECoG), which directly records electrical signals from the cortical surface, can successfully capture neural activity relevant to produced speech 2–9. However, speech-BCI advancements have largely focused on decoding a single language, primarily English or Dutch, due to study-population sampling 3,5–8,10–15.

This shared articulatory representation offers key advantages for multilingual BCIs. BCIs require training data with the user to develop high-performance decoders. This puts a burden on the user, and long training times may discourage adoption and continued use 64,65. Here, we demonstrate that, through transfer learning, training time for multilingual participants to use multiple languages may be strongly reduced. Further, our sub-lexical decoders achieved robust performance on syllables in a new language with no additional language-specific training data. These findings suggest that future multilingual speech BCIs may target shared sub-lexical units, for example phonemes 14, to improve bilingual vocabulary sizes and decoding speeds, by leveraging alternative approaches to fixed-window word decoding 8,15,66.

Although this is a single-participant study, the strength of shared cortical-articulatory representations is encouraging for generalizability to other patients. Notably, as this participant learned L2 (English) later in life, this may be even more encouraging for those who learn L2 early in life, which has been shown to lead to stronger shared representations 30,31,33,35. Future work, however, should directly investigate whether this effect varies with L2 proficiency, age-of-acquisition, and articulatory similarity to L1. While this study leveraged invasive electrophysiology, the results also hold relevance for non-invasive BCIs. A potential advantage of non-invasive BCIs is the ability to sample a larger distribution of cortex involved in functions beyond articulation. Non-invasive BCIs may apply a similar framework, as applied here with articulatory features, for other shared language features, such as semantics 67,68, that could enable rapid generalizability of a BCI across languages. These studies may also find language-specific signals, which we did not observe in speech-motor cortex, that may exist in higher-order cortical regions 28,69,70. A complementary direction for invasive BCIs is to study whether language-specific signals exist at the level of single neurons.

8) ***Line 32, "Together, these findings establish a multilingual cortical articulatory representation that persists 33 after paralysis and enables flexible decoding of multiple languages." — this is a proof of concept, N of 1, does not "establish"***

The reviewer raises a good point. We have changed the word “establish” in this sentence to “demonstrate” to better highlight this as a proof of concept study. We have also added a sentence highlighting the N=1 limitation in the discussion.

Together, these findings demonstrate a multilingual cortical articulatory representation that persists after paralysis and enables flexible decoding of multiple languages.

**9)** ***Line 65 -> words -> word***

We thank the reviewer for their careful reading of the manuscript and have corrected this typo.

To naturally decode bilingual phrases, it is desirable for the system to flexibly infer the intended language of the participant correctly based on cortical activity and natural language models, which capture language-specific word sequence statistics, alone.

**10)** ***Incorrect figure references? e.g. Fig. 2C on line 271?***

We thank the reviewer for catching this and have edited the manuscript to reference Fig. 2D.

Given evidence in the literature that L1 and L2 may activate distinct cortical regions, we next examined whether classification models trained only on English or Spanish words utilized different electrodes or features. We found that for both the HGA and LFS feature types, electrode contributions to the classifier (see Methods) were similar between English and Spanish (Fig. 2d).

**11)** ***“Confusability” should be defined / described in text***

This is another helpful comment from the reviewer. We agree that “Confusability” is not a standard term and should be defined in the results section. To this end, we have added a sentence explaining what we mean by confusability.

To assess this, we used a model trained on all 104 bilingual-words (no masking) and explored which factors drove confusability between any given pair of words (Supplementary figure S2). Here, confusability refers to the number of times that a given target word was incorrectly classified as another word in the dataset.

**12) The manuscript switches from % error to % accuracy. This is confusing, why not stick to one of the two metrics?**

This is a fair point raised by the reviewer. We chose to switch between these two terms based on the context of the modeling task. It is standard in speech processing fields (such as automatic speech recognition; ASR) to report decoding tasks in the form of percent error rates. More specifically, in ASR, transcriptions from audio are quantified by reporting the word error rate between the ground truth and transcriptions. The same standard has been used for BCI studies that decode neural activity into sentences [Willett et al. 2021, *Nature*; Moses et al. 2021, *NEJM*; Metzger et al. 2022, *Nat Comm;* Willett et al. 2023, *Nature*; Metzger et al. 2023, *Nature*, 2023], as done in this work. For this reason, we chose to report the % error rate in figure 1. For isolated classification tasks, on the other hand, the % accuracy is more commonly reported in studies that examine individual words [Moses et al. 2021, *NEJM*; Metzger et al. 2022, *Nat Comm;* Wandelt et al. 2022, *Neuron*; Berezutskaya et al. 2023, *J Neural Eng.;* Soroush et al. 2023, *Neuroimage*]. Given that our study combines approaches to both decode sentences from neural activity and offline analyses to classify neural activity as a single word, we switched between using the standard % error and % accuracy metrics. We would prefer to keep using both of these terms, to remain consistent with our and others prior work; however, if the reviewer strongly prefers we are open to standardizing across the manuscript.

**13) Lines 217-218 — The phrasing of this paragraph is confusing. I believe the "overall" condition is when the language is not conditioned during decoding?**

We thank the reviewer for catching this. We have changed "overall" to "all phrases" for consistency with earlier paragraphs. We have also changed the "overall" label in Fig. 1 to be "all phrases" for further consistency. "All phrases" refers to the fact that the reported metrics were computed across all of the evaluation (test-set) phrases, regardless of language. We hope these edits have added clarity to the reader.