nature portfolio

Peer Review File

An instantaneous voice synthesis neuroprosthesis

Corresponding Author: Dr Maitreyee Wairagkar

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Wairagkar et al present a real-time streaming speech synthesis system in a patient suffering from ALS. The group implanted 4 Utah arrays into the ventral motor cortex of the patient. They feed 600 ms of spike-band power and threshold crossing features into a Transformer Encoder to predict speech parameters that are subsequently transformed into an audio waveform by LPCNet. This whole process is very fast (10 ms delay) and allows for immediate voice feedback for the dysarthric patient. The reconstructed audio is of very high quality and also allows for modulation of pitch and accentuation, which gives the patient access to paralinguistic aspects of speech.

The presented pipeline is a well-designed prosthesis that presents a clear advantage for this patient. In particular, the generation of target speech, which the patient can't produce anymore, is clever. However, I have major concerns whether this prosthesis really relies on neural data:

The authors note that all four arrays show strongly increased activity for accentuation of words. The authors state: "A functional neuroanatomy result observed in this study that would not be predicted from prior ECoG and microstimulation studies is that the neural activity is correlated with paralinguistic features across all four microelectrode arrays, from ventral-most precentral gyrus to the middle precentral gyrus." (ll.377). To me, this pattern of activity is deeply concerning. In my understanding, it is highly unlikely that all electrodes on all four arrays in the motor cortex should show such a simultaneous strong increase in activity just because a word is accentuated (Fig 3C). The increase is almost two-fold for some arrays (Extended Data Fig 7). Particularly the fact that this finding has not yet been observed in microstimulation studies should make the authors cautious. In my opinion, a far more likely explanation for this increase is contamination through artifacts induced by the face and body movements of the participant. When watching video 11 carefully, it can be seen that the patient moves a lot more during the accentuated words. Such artifacts could easily explain the notable increase of activity on all four arrays. Such speech induced artifacts have been observed before [1-2]. Vibrations during mimed speech production (which is also a condition here) have even been used to reconstruct speech from a vibration sensor to a similarly high degree [3] as reported here. The vibration data collected in [3] also lies in the same frequency range as the extracted spike band power this work relies on.

To ensure that this prosthesis is not based on artifacts resulting from the patient's movements, I suggest attaching a vibration sensor (stethoscope or IMU) behind the ear of the patient and then recording during attempted speech production. This data can then be likened to the recordings from the Utah arrays and used to predict the LPCcoefficents independently. Without this control, it cannot be ruled out that the speech reconstruction in this study is based on artifacts.
On the other hand, this might actually be good news for patients. The results presented in this study and the NEJM study by the same group could benefit patients without the need for brain surgery.

Further points:
* Given the very high quality of the output, the intelligibility assessment is not well suited. Clearly, the six-choice sentence selection is far too easy, as the authors observed ceiling effects of almost 100% accuracy. I think an open transcription task on the reconstruction would provide a better assessment.

* A speaker identification task should be conducted (using listeners on Mechanical Turk, again) to assess whether the reconstructed voice (in the own-voice condition) really matches the participants old voice.

* I suggest to additionally cite the first paper that used a Transformer-Encoder architecture to reconstruct spectrograms from neural data [4] as the proposed pipeline is very similar.

* To put the MCD results of the acausal pipeline (l. 148) in perspective for the reader, it would be good to report the results of the causal one here, too.

[1] Bush, A., Chrabaszcz, A., Peterson, V., Saravanan, V., Dastolfo-Hromack, C., Lipski, W. J., & Richardson, R. M. (2022). Differentiation of speech-induced artifacts from physiological high gamma activity in intracranial recordings. NeuroImage, 250, 118962.
[2] Roussel, P., Le Godais, G., Bocquelet, F., Palma, M., Hongjie, J., Zhang, S., ... & Yvert, B. (2020). Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. Journal of Neural Engineering, 17(5), 056028.
[3] Shah, N., Sahipjohn, N., Tambrahalli, V., Subramanian, R., & Gandhi, V. (2024). StethoSpeech: Speech Generation Through a Clinical Stethoscope Attached to the Skin. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(3), 1-21.
[4] Shigemi, K., Komeiji, S., Mitsuhashi, T., Iimura, Y., Suzuki, H., Sugano, H., ... & Tanaka, T. (2023, June). Synthesizing speech from ecog with a combination of transformer-based encoder and neural vocoder. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Referee #2

(Remarks to the Author)
Summary of the key results
The manuscript by Wairagkar ea reports on realtime decoding of attempted or mimed speech by a person with severe paralysis due to ALS. Small chunks of Utah-array data (256 electrodes) were directly converted to audio with a delay of some 50 ms, and the result was promising in terms of intelligibility. Moreover, authors could identify changes in emphasis/intonation and pitch, which likely further enriches the users' experience of speaking via a computer.

Originality and significance: if not novel, please include reference
The results are very important in that the user can express him/herself vocally, at a speed that would allow participation in conversation with company. At this point it is not clear yet whether decoding will be feasible outside of research context, for instance with other peoples voices potentially degrading decodability of own speech. Moreover, it is not clear whether decoding is equally feasible in people who can not mime due to facial paralysis, such as in the case of late-stage ALS. Nevertheless, the results pose a significant advance in Brain-Computer Interface research in terms of speed and intelligibility. I do have some concerns, notably about the measure of intelligibility, amount of training required plus stability of decoders over time, correlated brain signal features for paralinguistic variables and generalizability of findings to other people with severe paralysis.

Data & methodology: validity of approach, quality of data, quality of presentation
The analytical approach is appropriate and straightforward, although I have concerns about how intelligibility was quantified. Data quality and presentation is quite adequate

Appropriate use of statistics and treatment of uncertainties
This is a single-subject study hence statistics are only within-subject. They are appropriate in terms of statistical evaluations of reported results of experimental tests

Conclusions: robustness, validity, reliability
Results seem robust, but there are some unclarified issues regarding training data, stability of decoders and generalizability to people with facial paralysis

Suggested improvements: experiments, data for possible revision
The measure for intelligibility is not quite adequate, it is suggested to apply a more rigid and quantifiable measure for each single word

References: appropriate credit to previous work?
References are adequate

Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions
The manuscript is written concisely and clearly. Conclusions may need additional information

Major comments

Regarding a key measure of performance, intelligibility, I am rather surprised at the approach the authors took. Intelligibility was assessed by identification by 15 volunteers ('We quantified intelligibility by asking 15 human listeners to match each of the 956 evaluation sentences with the correct transcript '(choosing from 6 possible sentences of the same length)'. This is a rather forgiving test since with only a few intelligible words (or even one if sentences differ on the first) a whole sentence can be identified out of 6. Yet authors then state that 'These results demonstrate that the real-time synthesized speech ... can be identified by non-expert listeners.' This seems a rather big leap in interpreting results. To get a better estimate of intelligibility I would prefer to see a test where volunteers do not choose between sentences but write down for every single word what they think they heard. Tis can then be compared to the prompted word and provides a much more informative measure. A

need for a better measure is further indicated by the conclusion in Discussion that 'The resulting voice was often intelligible' which contrasts with the near 100% performance on the 6-choice test.

Authors do not mention the fact that T15 used movement of his articulators to generate the brain signals for brain-to-speech decoding in the Discussion section. As I assume authors are aware of, somatosensory feedback signals are present in primary motor cortex (eg doi.org/10.1007/BF00237846, doi:10.1152/jn.00949.2014, doi.org/10.1101/2023.08.05.552025) and may well contribute to decoding of speech. Authors should discuss this and add to the Limitations section as a constraint to generalizability the potential ramifications for decodability of speech in people with facial paralysis or complete paralysis (eg LIS)

A frequent occurring challenge with speech decoding is training and calibration, an issue that is typically addressed in BCI publications. I am sure it is distributed across the methods section but can the authors clearly indicate the information readers need to assess how much time the participant spent generating data for model training, and stability of decoders over time. Include a timeline with days starting on the first day of generating training data and ending with the day of the last test, and make clear the time (days) between last decoder model training and the test.

Fig3: loudness, intonation and pitch were studied separately, and it seems they are all associated with amplitude of spike band power. Could they measure one and the same phenomenon (ie a single paralinguistic feature)? Can be decoded independently if co-occurring in attempted/mimed speech? please clarify

Minor comments

Line 52: imagining instead of imaging

General: please elaborate on what 'causal' and 'acausal' decoding mean other than (near) realtime. Perhaps those terms don't add anything to the discourse. Surely brain signals constitute cause to speech decoding even if not immediate?

Line 65 : ‘To overcome this, we generated synthetic target speech waveforms from the prompt text' Please explain what the prompt text refers to

Line 180: 'The accuracy of his free response synthesis was slightly lower than that of cued speech'. How was this determined, did the participant also produce or indicate the text he wanted to voice?

Fig 3: previous studies suggest that speed of speaking affects neural activity (eg doi: 10.1016/j.neuron.2016.01.032, www.nature.com/articles/s41583-020-0304-4), did the authors assess decoding performance differences for slow and fast speech? This could be important for a users' performance optimisation.

Referee #3

(Remarks to the Author)
The study reports an instantaneous voice synthesis neuroprosthesis. Previous studies have successfully decoded imagined speech from intracranial neural signals, but the output is generally text. The current study, however, decodes the speech sound and therefore advances previous studies in two aspects: (1) decoding the timing of syllables and (2) decoding pitch information that is not present in the text. I think the two advances are of potential scientifical importance and are critical for potential application of the method in ALS patients. However, I have major concerns about what signals are actually decoded and whether the results are appropriately evaluated.

1. Decoding of syllable timing: From the videos, it is clear that the patient moves whenever trying to produce a word. On the positive side, I wonder if this signal could be combined with the BCI to further improve performance. For example, I wonder if the syllable decoder can be initially trained based on body movements, i.e., treating the onset of each stereotypical body movement as the onset of a word when training the syllable decoder. The authors may also consider to directly decode syllable boundaries based on the video and replace the neural syllable decoder with the video syllable decoder to see which achieves better speech synthesis performance.

On a slightly negative side, I wonder if the body movements are driving the ‘null space’ activity. In other words, the ‘null space’ activity is not related to speech per se but is related to the overt body movement, which is absent or much weaker in heathy individuals. Furthermore, I wonder the syllable decoder could actually read out signals controlling such overt body movements, instead of the covert syllable boundaries. It’s absolutely fine in terms of the application but scientifically I wonder if the authors could clarify what kind of signal contributes to syllable boundary decoding. I also wonder if some channels behave like the output-potent neural activity.

2. Speech rate: A sharp difference between the decoded speech and normal speech is the speed. Normal people speak 5-8 syllable per second but the decoded speech is at most 1-2 syllable per second, with long pauses between words. I don’t think it’s a problem but it needs to be discussed – Why did the patient speak so slowly and lose the ability to speak fluently? Did the patient try to slow down to cooperate with the BCI or which part of the brain might have been affected by the disease so that the person can no longer attempt to speak fluently?

3. Results evaluation: The correlation coefficient between target and synthesized speech is used to evaluate speech decoding, but this measure is not particularly informative. First, the high correlation is driven by the long pauses in patient

speech. Even if the synthesized speech decodes a completely different set of words, it will have high correlation with the target speech as long as the words are synchronized to the target speech. Second, the synthesized speech naturally synchronize with the target speech since the timing of syllables are decoded from the same neural signals for both target and synthesized speech, although in different manners. Furthermore, the paper suggests that target speech is synthesized based on the text shown to the patient. However, target speech is also present and is used for evaluation for spontaneous speech. I wonder how the target speech is constructed in these scenarios. If the target speech is synthesized online during speech production (through instantaneous text decoding and instantaneous TTS), it is an alternative method for speech synthesis.

Compared with the correlation with target speech, open-set speech recognition is a much more reasonable evaluation method. It can be done either with automatic speech recognizers (ASR) or with human evaluators. Even human evaluation is not costly, a few naïve listeners will be enough. Such a test is critical if the purpose of the work is to enable the patient to interact with others. Similarly, the performance of pitch decoding should be judged by humans (e.g., taking out a pair of syllables from the melody condition and asking human participants to judge which one has a higher pitch). Based on the videos, the melody seems to be quite difficult to judge for myself so that I would like to see ratings from a group of naïve listeners.

Minor issues:
Terminology: The summary says the study decodes “prosody and intonation”, and the discussion says the study decodes “pitch and intonation”. Notably, pitch, prosody, and intonation are not different things, and it is confusing to list them together. For example, it’s not that the study decodes pitch and something else called intonation that is not based on pitch. Pitch and loudness are critical dimensions for prosody and are decoded in the study. However, the two features are separately decoded in different experiments. As far as I can tell, they are not integrated to form “prosody”. It’s better to say that the study can decode pitch, loudness, and speech rate.

Loudness decoding: The study decoded the loudness of an utterance, and I think a few more analyses are necessary to establish what is decoded is loudness. First, I wonder if the same decoder could determine that the mimed speech is softer or quiet or whether the emphasized words are louder. Second, please show the instantaneous output of the decoder when applied to the imagined melodies, questions, and statements like in Fig. 3. I wonder if aligns with the pitch decoder. If loudness decoding is a part of the conclusion, it should be strengthened. However, I don’t think removing this part undermines the study.

What is MCD? It’s not explained or spelled out.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
I am very impressed by the amount of work and rigor the authors invested into addressing all reviewers' comments. I am a lot more convinced now that the impressive results are indeed based on neural activity. Furthermore, the new intelligibility analysis very realistically reflects the acoustic output. Congratulations on this impressive work.

Despite the higher WER than in the "to-text" approach, I believe this project to represent an important step forward for speech neuroprostheses. By providing immediate feedback and allowing the patient to use intonation and accentuation, the manuscript brings the full expressive power closer to patients.

Methods are well described and statistics used appropriately. The relevant work is credited and the new findings placed in the existing literature.

The only thing I think is necessary before this work can be published would be to make the new, parallel recording of stethoscope, IMU and Neural data with the appropriate prompts available to the general public so that the fusion and each modality individually can be investigated by the community.

Referee #2

(Remarks to the Author)
I thank the authors for their careful and thoughtful address of my concerns. I am satisfied with the revisions.

one small confusion remains: regarding the new intelligibility test: you mention 23 female plus 5 male naive listeners but metrics are reported for 7. Perhaps not all 28 did their evaluations in person? or more likely 23 is a typo

Referee #3

(Remarks to the Author)
The authors have done a good job revising the manuscript, and have incorporated additional data that can help us to

interpret what signals are read out by the speech BCI. I'm largely satisfied with the revision, but two points warrant further discussion in the paper.

1. Regarding the speech rate, I fully appreciate that the patient experiences significant difficulty in moving his articulatory muscles. However, in principle, the performance of a BCI should not be constrained by muscular control capabilities, as the signal is directly decoded from brain activity. Therefore, the authors should still discuss why the patient speaks slowly even when it is not necessary to control articulatory muscles: For example, in the future, is it possible to design a speech BCI that does not depend on the patient's ability to overtly articulate? Or, does the patient have difficulty 'imagining' fluent speaking?

2. Output-potent/null space: The output-potent/null space is defined as neural dimensions that can/cannot linearly decode speech features. It is then demonstrated that both spaces can support nonlinear speech decoding. Thus, the distinction between output-potent/null spaces appears to reduce to whether a neural dimension is linearly or nonlinearly related to speech features, which may not be fundamental difference between output-potent and null spaces.

Response to Reviewers

Title: An instantaneous voice synthesis neuroprosthesis
(Manuscript ID: 2024-09-19838)

Dear Editor and Reviewers,

We thank all the reviewers for their insightful and detailed feedback on our manuscript. We appreciate that all three reviewers expressed enthusiasm for this study (qualified with some technical questions, which we address below): "*The reconstructed audio is of very high quality and also allows for modulation of pitch and accentuation*" (Reviewer 1); "*The results are very important in that the user can express him/herself vocally, at a speed that would allow participation in conversation with company… the results pose a significant advance in Brain-Computer Interface research in terms of speed and intelligibility*" (Reviewer 2); "*I think the two advances are of potential scientifical importance and are critical for potential application of the method in ALS patients*." (Reviewer 3).

The reviewers' questions and comments have helped us to make this manuscript stronger and more thoroughly demonstrate the robustness of the brain-to-voice BCI approach. At the beginning of this Response to Reviewers, we first address the two main concerns that were shared by multiple reviewers and also highlighted by the Editor. Later in this document we address each reviewer's other individual comments in detail. Throughout this document, our response to the reviewers is in **blue** text like this**.**

> Relevant copied passages from the manuscript are independent in **black** text, like this.
> The new text added to the manuscript is highlighted in **red**.

ADDRESSING REVIEWERS' TWO ESSENTIAL QUESTIONS

The two main concerns raised by the reviewers as summarized by the editor were: (1) the need for a more sensitive intelligibility measure for evaluating synthesized voice and (2) the possibility that retained movement by the participant caused motion artifacts in the neural recordings that contributed to voice synthesis quality. We have performed multiple additional analyses and collected new experimental data to address these comments as follows:

1. Quality of performance assessment of synthesized speech

Open transcription of synthesized speech by naïve human listeners

We agree with the reviewers that the manuscript would be strengthened with a measure of intelligibility by human listeners that is more sensitive than the multiple-choice selection we originally included. In particular, we recognize and agree that the almost-at-ceiling accuracy of the multiple-choice selection made it a sub-optimal metric, and that an evaluation scheme more similar to the ultimate goal (intelligibility, not classification) is desirable. To obtain this more nuanced metric of intelligibility, we recruited naïve listeners to perform open transcription of synthesized voice trials.

We randomly selected a subset of 90 voice synthesis trials from our closed-loop evaluation set (spanning 8 sessions) and asked 7 naïve listeners to transcribe each sentence. Unlike our multiple-choice listening assessment, which used online gig workers (consistent with prior

studies in the speech BCI field), here we recruited members of the local community and ran the assessments in-person, which we believe is closer to a “gold standard” human listener evaluation (albeit one that is more expensive and labor-intensive). Additionally, the open transcription was also conducted on a set of 38 trials included in the main videos 2, 3 and 4 so that readers can get a quantitative assessment of the intelligibility of the examples they hear and can contextualize their own impression of these examples’ intelligibility with the reported metrics. We believe that this is an important thing to do given concerns in the field that some of the audio-video examples from past studies may be extra “cherry-picked” and not representative of average results. Finally, to compare the intelligibility of the BCI-synthesized speech to participant T15’s residual dysarthric speech, we also asked the human listeners to transcribe 20 sentences of his attempted speech (as recorded during the BCI research session trials between post-implant days 172 to 195).

For the general synthesis evaluation set, we obtained a median phoneme error rate (PER) of 34.00% and median word error rate (WER) of **43.75%**. For the video examples the median PER was 23.73% and median WER was **37.50%**. To our knowledge, this is the lowest error rate achieved by a closed-loop synthesis BCI. For T15’s dysarthric speech, the median PER was 83.87% and median WER was 96.43%. This indicates that T15’s residual speech was almost completely unintelligible to the naïve listeners whereas the BCI synthesized voice vastly improved the intelligibility. We were pleased to discover that most words were correctly understood by naïve listeners, and believe that these metrics plant a flag in the sand for the state-of-the-art and are an encouraging step towards consistent intelligibility for individuals with severe dysarthria. It is important to note that the sentences synthesized in this study were chosen at random and lacked context, whereas in a real-world conversation scenario where there is context across sentences, intelligibility is likely to be higher for most listeners. Furthermore, learning to understand the user’s synthesized voice with more exposure may also increase intelligibility.

We have added this new intelligibility measure in our main results **line 141** and the main **Fig. 2l.** Details of the process are outlined in Methods **line 777**.

Fig 2. l line 197:

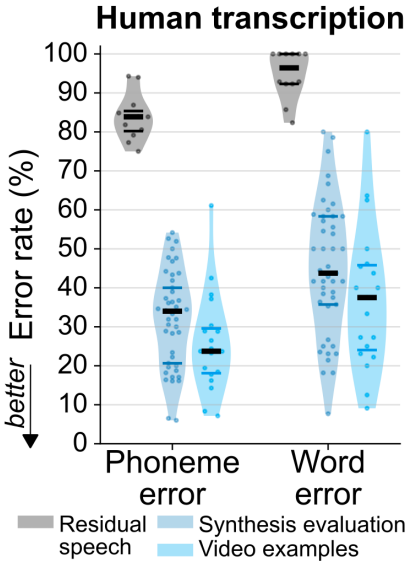


Fig. 2. Voice neuroprosthesis allows a wide range of vocalizations. … **l.** (Right) Open transcription of BCI synthesized speech by human listeners. Phoneme and word error

rates of the BCI synthesized speech (blue) and T15's residual dysarthric speech (black) are shown. The bold black line shows the median error rate, and thin colored lines show the (bottom) 25th percentile and (top) 75th percentile.

Results line 141:

Second, we conducted open transcription tests by naïve human listeners (**Fig. 2l right**). The transcription median phoneme error rate (PER) was 34.00% and median word error rate (WER) was 43.75%. In contrast, T15's residual dysarthric speech was transcribed by the same human listeners with median PER of 83.87% and WER of 96.43%. This indicates that the brain-to-voice BCI vastly improved T15's intelligibility.

Methods line 777:

Human evaluation of synthesized speech – assessment 2: open transcription. We also conducted an open transcription task to obtain a more sensitive metric of intelligibility of synthesized speech, especially considering that the transcript matching accuracies were close to ceiling effects. We recruited naïve listeners from the UC Davis community (native English speakers, 23 female, 5 male, aged 18-42 years) to transcribe each sentence from a representative subset of 90 randomly selected voice synthesis trials from the evaluation set. By "open" transcription, we mean that the evaluators were instructed that the sentences they would hear could use any English words (there was no limited vocabulary for them to choose from). For quality control, human evaluators were invited in person to perform the transcription task using headphones and a computer in a quiet space. The evaluators were instructed to type out what they heard, as best they could, after listening to each sentence (which they could repeat multiple times if needed). Participants were remunerated for performing the evaluation task for up to an hour. To familiarize these naïve listeners with the BCI voice and T15's pace of speaking (which is much slower than healthy speech), a short 5-minute video was shown which played several audio examples of synthesized voice and displayed their corresponding transcripts as part of their brief orientation before they started the task. The evaluators reported that this video helped them to understand the transcription task better and know what to expect, but that this brief orientation was not sufficient to learn to understand the voice. Each of the 90 evaluation set trials was transcribed by 7 human evaluators. Additionally, open transcription was also conducted on 38 trials from main **Videos 2, 3** and **4** to help readers quantitatively assess these examples and contextualize what they hear with the overall metrics. To compare the intelligibility of the BCI synthesized speech with T15's residual dysarthric speech, we also asked the human evaluators to transcribe 20 trials of T15's attempted speech recorded via a microphone during the research blocks.

The resulting transcriptions were quantified by computing phoneme error rate (PER) and word error rate (WER) – i.e. percentage of phonemes or words transcribed incorrectly. To avoid quantization and variability introduced by very short sentences (e.g., all four-word sentences could only yield WERs of 0%, 25%, 50%, 75%, or 100%), we grouped two sentences together into longer sentence-pairs with a large range of word lengths before computing error rates. For each sentence-pair, median PER and WER from all 7 transcribers was computed.

2. Absence of motion artifacts in intracortical neural recordings

The second main reviewer concern was to ask whether some of the neural decoding performance could actually be due to the participant's retained (but limited) ability to move and make dysarthric vocalizations. Intracortical recordings are generally known to be resistant to external motion artifacts or muscle artifacts since these implanted devices record neural signals directly from the source - single neurons in contrast to non-invasive neural signal recording methods such as surface EEG that are more sensitive to muscle or vibration artifacts. There have been several major recent speech BCI studies using implanted electrodes where the participant could still move but the speech decoding was not attributed to these residual movements including: Willett et al. "A high-performance speech neuroprosthesis" *Nature* 2023 [participant T12 could still speak somewhat intelligibly in addition to moving her head and hand]; Metzger et al. "A high-performance neuroprosthesis for speech decoding and avatar control" *Nature* 2023 [participant BRAVO-3 is anarthric but can still move her head]; and our own previous report about T15's "brain-to-text" decoding performance, Card et al. "An Accurate and Rapidly Calibrating Speech Neuroprosthesis" *New England Journal of Medicine* 2024.

That said, we did not take this for granted because of other studies and assumptions in the field. Rather, to provide additional confidence that artifacts did not affect accuracy in this particular study, we have now conducted multiple additional analyses to show that the neural recordings were not contaminated by movements during attempted speech and that T15's residual speech did not contribute to speech decoding and voice synthesis. Specifically, we:

1. Performed an acoustic contamination analysis
2. Showed variety in speech tuning across different electrodes
3. Demonstrated that the brain-to-voice decoder does not output speech during non-speech vocalizations
4. Recorded muscle vibrations and motion during attempted speech using additional surface sensors (stethoscope and IMU) and attempted to decode speech from these biosignals
5. Performed a video analysis to track T15's head movements during attempted speech

Through these analyses, we demonstrate that despite T15's retention of limited orofacial movement and the ability to vocalize unintelligibly, the BCI reported in this manuscript utilized neural correlates of speech to synthesize voice in closed loop and was not contaminated with motion or acoustic artifacts.

2.1. Acoustic contamination analysis

We performed an acoustic contamination analysis on neural recordings during speech synthesis trials to rule out the possibility of acoustic contamination of neural signals due to T15's vocalizations during attempted speech using the now-gold-standard method described in *Roussel et al., 2020* [ref 1*]. This method quantifies any possible acoustic contamination statistically by computing correlations between frequencies in the spectrum of the audio and the spectrum of each electrode.

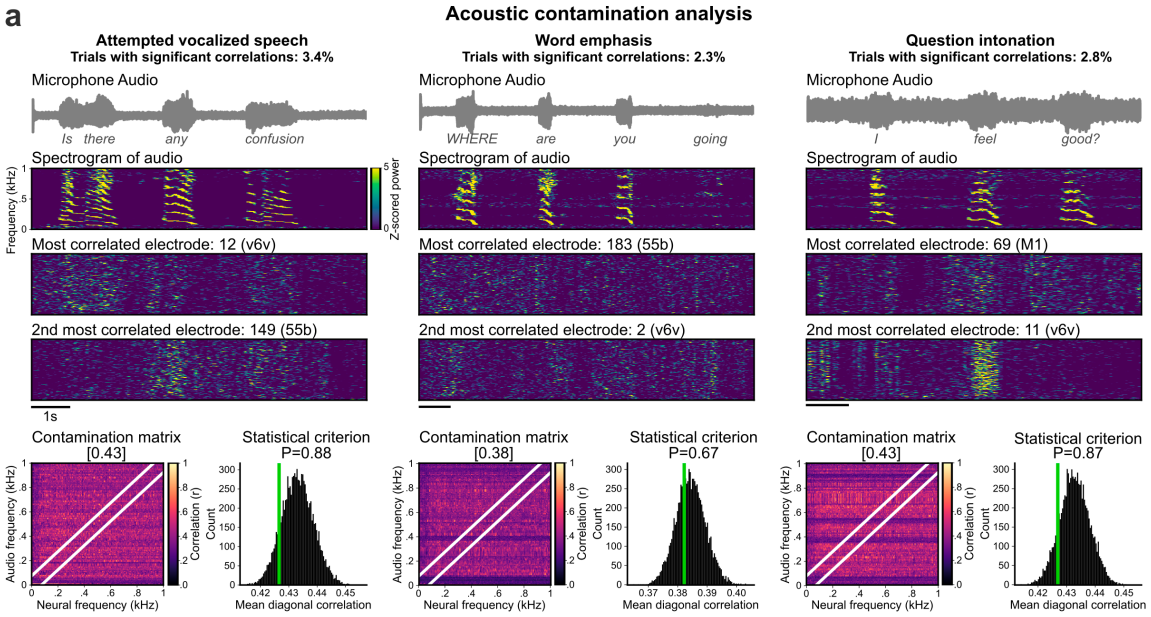
We assessed a subset of representative neural trials from all weeks of evaluation data collection and included the three main types of vocalized speech tasks used for closed-loop speech synthesis: (1) *regular attempted vocalized speech*, (2) *word emphasis*, and (3) *question intonation modulation*. This analysis showed that only 3% of trials had any statistically significant ($p<0.05$) evidence of *potential* acoustic contamination (in line with what is expected by chance

and is considered evidence of “clean” recordings) and which cannot plausibly have substantially affected the performance of closed-loop voice synthesis.

[1*] Roussel, P., Le Godais, G., Bocquelet, F., et al., (2020). Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. Journal of Neural Engineering, 17(5), 056028.

In the revised manuscript, the acoustic contamination analysis results are now shown in **Extended Data Figure 6a**. We have described the methods in the Methods **line 912** and results in the Results **line 159** as follows:

Extended Data Figure 6a line 1074:



Extended Data Fig. 6. Neural activity is not contaminated by acoustic artifacts and residual vocalization and movement cannot synthesize intelligible speech. a. Three example trials’ audio recording, audio spectrogram, and the spectrograms of the two most acoustic-correlated neural electrodes. Examples are shown for the three types of speech tasks. The prominent spectral structures in the audio spectrogram cannot be observed even in the top two most correlated neural electrodes. An increase in neural activity can be observed before speech onset for each word, reflecting speech preparatory activity and further arguing against acoustic contamination. Note that in the word emphasis example, the last word ‘going’ is not vocalized fully (there is minimal activity in its audio spectrum), yet an increase in neural activity can be observed that is similar to other words. Contamination matrices and statistical criteria are shown in the bottom row, where p-value indicates whether the trial is significantly acoustically contaminated or not.

Results line 159:

We verified that the neural data were not acoustically contaminated by T15’s residual speech (**Extended Data Fig. 6a**) …

Methods line 912:

Control experiments to confirm validity of neural recordings

Acoustic contamination analysis

We performed an acoustic contamination analysis following the method described in [42] to test for the possibility of contamination in neural recordings due to T15's (unintelligible) vocalizations during attempted speech. This analysis statistically determined the likelihood of contamination by evaluating correlations between the spectrum of T15's vocalizations recorded by a microphone and the spectrum of each intracortical electrode's voltage time series.

We analyzed 627 representative trials sampled from all weeks of evaluation data from all the three types of vocalized speech tasks used for closed-loop voice synthesis: (1) regular attempted vocalized speech, (2) word emphasis, and (3) question intonation modulation. Some trials showed spurious acoustic correlations between microphone static and the background neural activity during silent periods, which were eliminated by removing background static from microphone recordings (by replacing the static only in non-speech silent sections of the microphone audio by zeros). The analysis resulted in 3.03% of trials identified as acoustically correlated (**Extended Data Fig. 6a**). This indicated that the majority of the speech synthesis trials were not acoustically contaminated, and the small percentage of correlated trials is consistent with chance and extremely unlikely to meaningfully impact closed-loop voice synthesis.

2.2. Diverse speech tuning across different electrodes

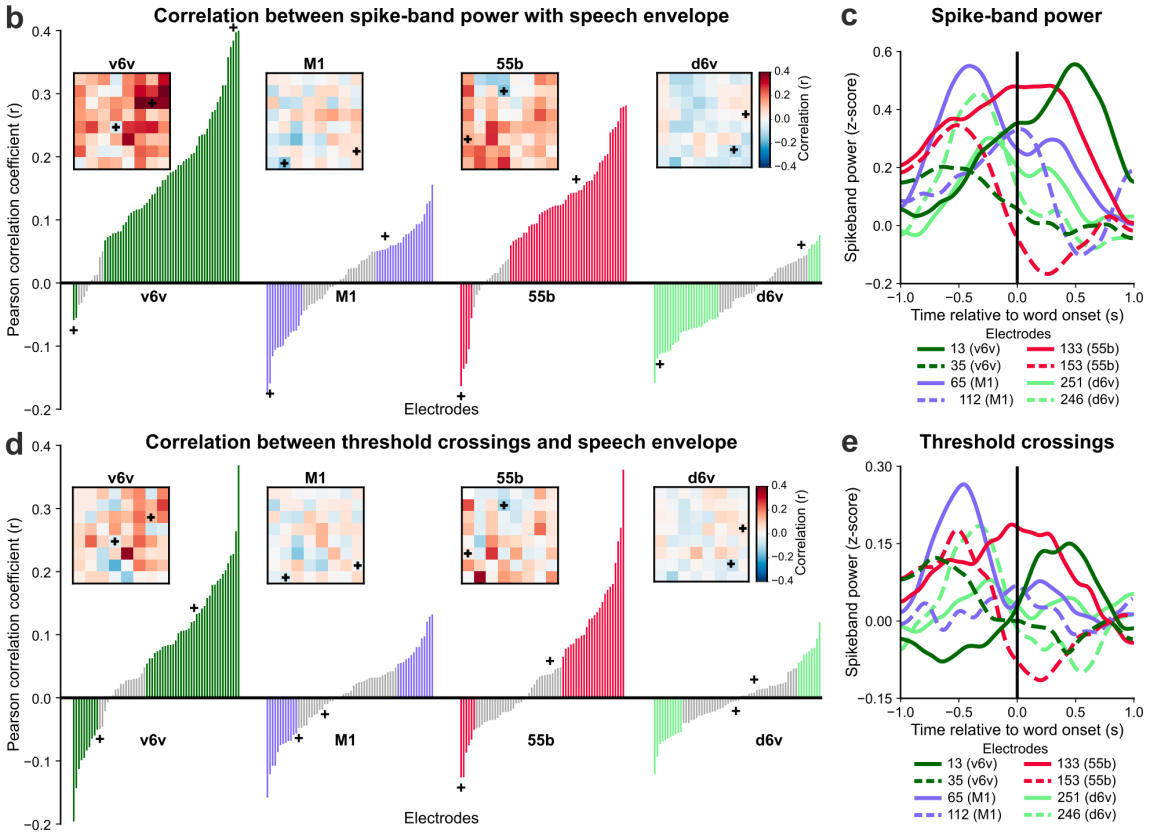
If the measured intracortical signals were artifacts related to the participant's retained vocalization or other physical movements, then we would expect that the voltage time series would look similar across all electrodes (after all, these vibrations would pass through all the arrays, or across both pedestals affixed to the skull). In contrast, neural activity from neurons near each electrode would exhibit a diversity of different modulation patterns and correlations with the attempted speech (e.g., some neurons would increase their firing rates, others decrease their firing rates, and at different times). To demonstrate that the intracortical recordings captured speech-related neural correlates with rich and heterogenous neural dynamics, we measured how each electrode was tuned to speech by computing the correlation between the (putatively) neural activity recorded on each electrode and the predicted speech envelope (**Extended Data Fig. 4b-e**). As expected from our results described in the original manuscript, we observed higher correlations in arrays v6v and 55b (which also show higher overall modulation in neural activity during speech), followed by arrays M1 and then d6v (our "least speechy" array). Each array contained a variety of speech tuning correlation coefficients, including both positively and negatively tuned electrodes, which is not consistent with artifacts.

Extended Data Fig. 4c,e shows the time courses of the trial-averaged neural activity of two example electrodes from each array (one positively and one negatively tuned to speech). Each electrode shows different time courses with respect to speech onset: some of the electrodes show higher activity before speech onset (possibly contributing to speech preparation) while other electrodes show increased activity during execution of speech or at the onset of speech. This provides a snapshot of complex neural dynamics across different electrodes in different arrays. This variety in neural dynamics rules out the possibility that the majority of modulation

reflects motion or acoustic artifacts in the neural signals. Additionally, arrays connected to the same percutaneous pedestal (v6v and M1 are connected to pedestal 1 and 55b and d6v are connected to pedestal 2) also show different speech tuning, which further confirms that there are no mechanical artifacts introduced by individual pedestals.

We have added the following tuning figure in **Extended Data Fig. 4** and refer to this in the Results in **line 154**.

Extended Data Fig 4b-e line 1041:



Extended Data Fig. 4. Electrodes show variability in speech tuning. …
b, d. Pearson correlation coefficients of spike-band power and threshold crossings, respectively, between each electrode and the speech envelope (first LPCNet feature predicted by the brain-to-voice decoder). Electrodes are grouped by array and sorted in ascending order to show that different electrodes have different tuning (both positive and negative) with speech. Arrays v6v and 55b have higher correlations with speech and the majority of electrodes show positive tuning. Arrays M1 and d6v have lower correlations with more electrodes tuned negatively. Electrodes with non-significant correlation ($p>0.05$) are shown in grey. Insets show the same correlations for each electrode arranged spatially in the array. **c, e.** The time course of average spike-band power and threshold crossings across trials of two example electrodes with positive (solid line) and negative (dashed line) correlations from each array (example electrodes are marked by the '+' symbol in (**b**, **d**)). Different electrodes show complex neural dynamics with respect to speech onset and a rich variety in speech tuning. For example, some electrodes have higher activity before speech onset (possibly contributing to speech preparation) while others have higher activity during or after speech onset.

Results line 154:

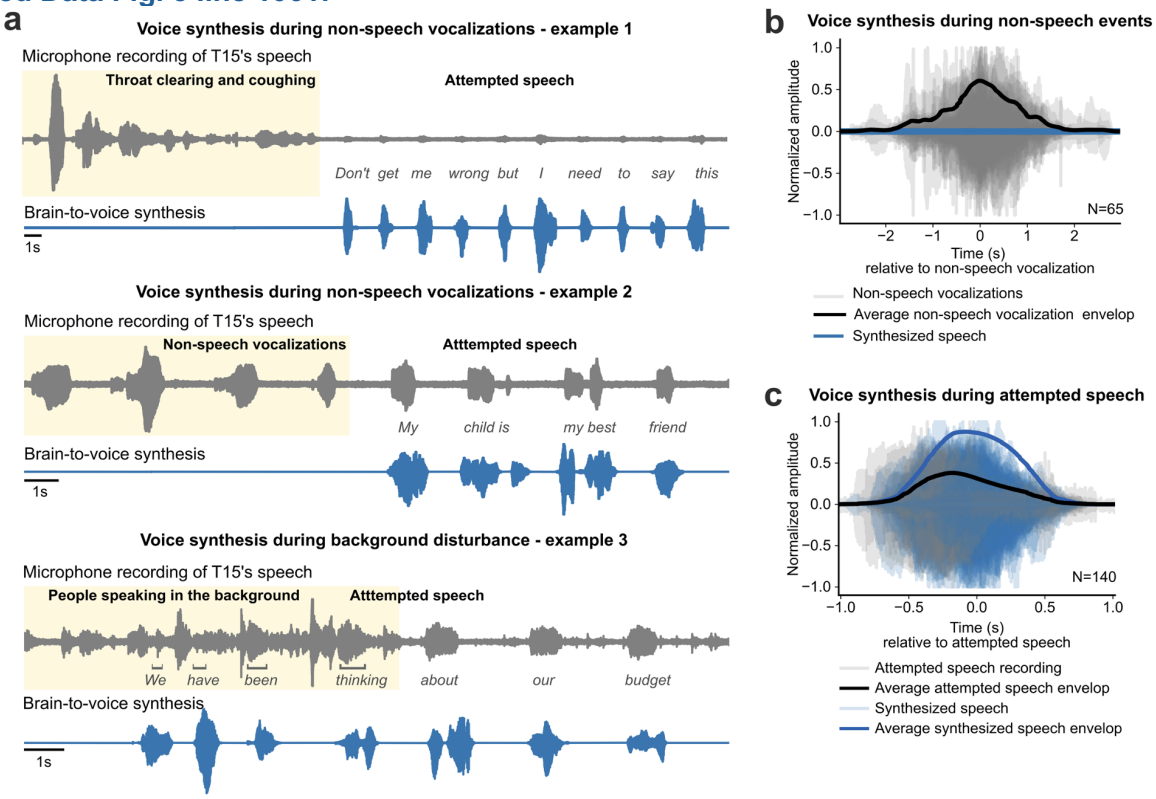
All four arrays showed significant speech-related modulation and contributed to voice synthesis, with the most speech-related modulation on the v6v and 55b arrays and much less speech-related modulation on the d6v array (**Extended Data Fig. 4**). Different electrodes showed a rich variety in speech tuning (**Extended Data Fig. 4b-e**).

2.3. The BCI does not output speech during non-speech vocalizations

One requirement of a speech BCI – along with high accuracy speech synthesis – is that it should not spuriously synthesize speech when the user is not voluntarily attempting to speak. If the neural recordings were impacted by movement or vibrations, then we would reasonably expect that non-speech vocalizations, like coughing, would also produce synthesized voice. We found that the brain-to-voice BCI showed very robust speech detection without false positives and the decoder synthesized voice only when T15 voluntarily attempted to speak. We have now shown this in **Extended Data Fig. 5** with examples of trials that illustrate that no speech was synthesized during orofacial movements such as coughing, clearing his throat, yawning etc. It also did not output voice in response to people speaking in the background. **Extended Data Fig. 5b** also shows several examples of non-speech vocalization events and the corresponding speech synthesis (which output silence) and similarly, **Extended Data Fig. 5c** shows samples of speech synthesized appropriately during attempted speech. Thus, neural activity during attempted speech is different from the neural activity during non-speech vocalizations or orofacial movements and the brain-to-voice decoder has implicitly learned to ignore these events and to avoid spurious unintentional speech synthesis. This further adds to the strong evidence that the neural data was not contaminated by the orofacial movements or vocalizations and that the voice synthesis was not affected by these factors.

We have added the following figure in **Extended Data Fig. 5** in the manuscript and cited this in the Results in **line 157**.

Extended Data Fig. 5 line 1061:



Extended Data Fig. 5. Speech is not synthesized during non-speech vocalizations or orofacial movements. a. Three example trials show microphone recording (grey) of T15 during attempted speech trials with coughing, throat clearing, non-speech vocalizations or people speaking in the background and the corresponding brain-to-voice synthesis output (blue). The brain-to-voice decoder did not synthesize audible speech and instead output silence during these non-speech vocalizations or when other people were speaking simultaneously (note that T15 starts speaking via the neuroprosthesis midway through the background conversation in example 3). In contrast, it did synthesize voice when T15 voluntarily attempted to speak. Speech was synthesized instantaneously at the exact pace with which T15 attempted to speak and with very low latency. **b.** Samples of non-speech vocalization events (grey) and the corresponding speech synthesis output, which was zero (silence) throughout all these events (blue). **c.** Examples of speech synthesis during attempted speech of each word. Here, the speech is synthesized (blue) appropriately as expected during attempted speech (grey).

Results: line 157

Brain-to-voice synthesis was robust to non-speech vocalizations such as coughing, throat clearing, yawning, or people talking in the background (**Extended Data Fig. 5**) and did not show spurious synthesis during these events.

2.4. Decoding speech from non-invasive biosignals

Here we address reviewers' concern that muscle or acoustic artifacts may have been introduced in the neural data during T15's attempted speech that could help voice synthesis. This is indeed a relevant concern especially given the recent interest in the literature in silent speech decoding from such non-invasive signals. To test whether T15's residual severely dysarthric speech (or related vibrations during attempted speech) could contribute to voice synthesis, we conducted a new control experiment where we recorded additional non-invasive surface sensors to record the vibrations/muscle movements and acoustics of his attempted speech. We recorded several biosignals simultaneously with neural signals by placing a stethoscopic mic on T15's left mastoid, an inertial measurement unit (IMU) sensor on T15's right mastoid and a microphone in front of T15 (as shown in **Extended Data Fig. 6b**). Our goal was to identify whether these signals could be used to accurately synthesize voice (if yes, this could indicate that our main decoding results could have benefited from contamination of neural data with these artifacts).

We trained separate decoders for each modality of recordings and synthesized voice from these biosignals. We observed that Pearson correlation coefficients comparing the synthesized voice with target speech (after eliminating long pauses between words) were significantly higher for speech synthesized from neural activity (0.87±0.05) than the speech synthesized from biosignals (0.74±0.05), with largely non-overlapping distributions [note that correlation coefficients are a poor proxy for intelligibility, so these numbers understate the actual difference in voice synthesis quality; this can be appreciated from the new video and the human transcription results described below]. We now show example audio trials of speech synthesized from each of these biosignals in comparison with the speech synthesized from neural signals in the new **Video 15** (https://ucdavis.box.com/s/tz4odg4uefo7ifsilxm92odv1lm5sj03) included in the revised manuscript. We compared the intelligibility of speech synthesized using neural decoding with the speech synthesized from stethoscopic microphone (which had highest signal-to-noise ratio among biosignals) by recruiting naïve human listeners (via Amazon Mechanical Turk) to perform open-vocabulary transcriptions of the synthesized speech trials. We obtained a median word error rate of 43.6% for neural synthesis, consistent with the main results above. In stark contrast, the word error rate was 100% for stethoscope signal-driven synthesis, indicating that speech synthesis from biosignals was completely unintelligible.

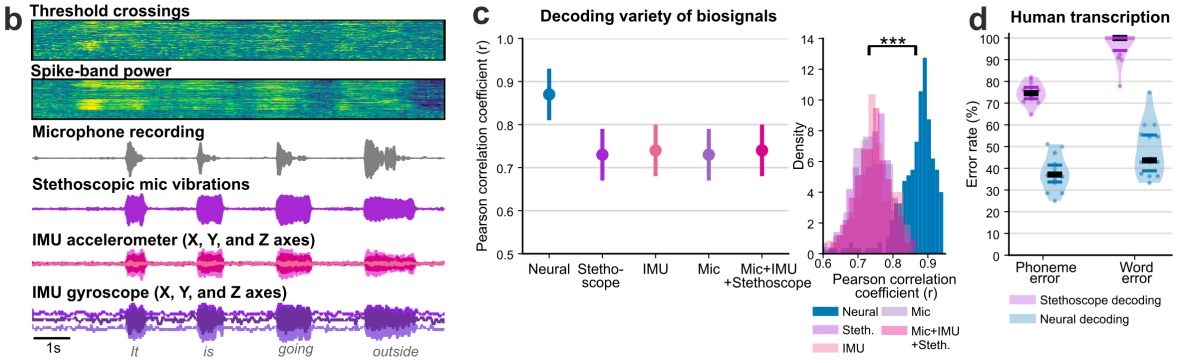
From these results we can conclude that acoustic and muscle vibration signals from T15's residual dysarthric speech did not contain sufficient speech-related information to synthesize intelligible voice and consequently even if they did bleed into the neural recordings (which the previously described results indicate was not the case), this would not help decoding as compared to the neural signal. That is, the brain-to-voice decoder that produced mostly intelligible speech decoded neural correlates of speech and not the artifacts of residual speech.

We have added the following details in the manuscript to clarify this.

Results: line 160

> … intelligible speech could not be synthesized from his vocalizations or vibrations from his residual movements (**Extended Data Fig. 6b-d, Video 15**).

Extended Data Fig. 6b-d: line 1074



Extended Data Fig. 6. Neural activity is not contaminated by acoustic artifacts and residual vocalization and movement cannot synthesize intelligible speech. …
b. An example trial of attempted speech with simultaneous recording of intracortical neural signals and various biosignals measured via a microphone, stethoscopic microphone and IMU sensors (accelerometer and gyroscope). Separate independent decoders were trained to synthesize speech using each of the biosignals (or all three together). **c.** Intelligible speech could not be synthesized from biosignals measuring sound, movement, and vibrations during attempted speech. (Left) Cross-validated Pearson correlation coefficients (compared to target speech) of speech synthesized using neural signals, each of the biosignals, and all biosignals together. Reconstruction accuracy is significantly lower for decoding speech from biosignals as compared to neural activity (Wilcoxon rank-sum, $p=10^{-59}$, n=240). (Right) Distribution of Pearson correlation coefficients of speech decoding from biosignals and neural signals are mostly non-overlapping, indicating that synthesis quality from biosignals is much lower than that of neural signals. **d.** To assess the intelligibility of voice synthesis from neural activity and biosignals (stethoscopic mic decoder), naïve human listeners performed open transcription of (the same) 30 synthesized trials using both the decoders. Median phoneme error rates and word error rates for neural decoding were significantly lower (43.60%) than decoding stethoscope recordings, which had word error rate of 100%. This indicates that intelligible speech cannot be decoded from these non-neural biosignals.

Methods: line 931

Decoding speech from non-invasive biosignals

We tested whether T15's residual speech or vibrations generated due to his vocalization carried decodable information, which could possibly introduce artifacts in neural recordings that would inflate decoding accuracy. We performed control experiments by simultaneously recording intracortical signals and speech-related biosignals non-invasively using a microphone to capture T15's acoustics; a stethoscopic microphone[43] (an acoustic stethoscope with a digital condenser microphone inserted in the stethoscope tube after removing the earpieces) placed on T15's left mastoid; and an inertial measurement unit (IMU) sensor (MPU-6050, HiLetgo) placed on his right mastoid to capture vibrations as T15 attempted to speak 240 sentences from a limited 50-word vocabulary (**Extended Data Fig. 6b**).

We trained separate decoders (using the same decoder architecture as the "brain-to-voice" decoder described above) for each modality of recordings (and a combined decoder for all biosignals together) using 10-fold cross-validation to synthesize voice offline (**Video 15**). We extracted 40 Mel-Frequency Cepstral Coefficients (sufficient to reconstruct voice) as input features from signals from the microphone and stethoscope, and from each of the 6 axes of the IMU sensor for the respective decoders. Thus, the microphone and stethoscope decoders had 40 input features, the IMU decoder had 240 input features, the combined biosignals decoder had 320 input features, and the neural decoder had 512 input features. A feature bin size of 10 ms, input sequence length of 600 ms, and all other decoder parameters were kept the same as neural decoding as described in the "brain-to-voice" decoder Methods sections above.

We computed Pearson correlation of voice synthesized from each of the biosignals with the target speech (**Extended Data Fig. 6c**), and also measured intelligibility of speech synthesized from neural activity and from stethoscope vibrations by recruiting human listeners via Amazon Mechanical Turk to perform open transcription of (the same) 30 randomly selected trials synthesized by the two decoders: each trial was transcribed by 7 human evaluators (using the same human listener evaluation method as described above). We used speech decoded from stethoscope vibrations as representative speech synthesis from biosignals because it showed the highest signal-to-noise ratio among all the recorded non-invasive biosignals. The median phoneme error rate (PER) for neural voice synthesis was 37.1% and the word error rate (WER) was 43.60% (consistent with human transcription results in **Fig. 2l**). In contrast, the median PER for biosignal decoding was 74.5% and the WER was 100% (**Extended Data Fig. 6d**). Thus, intelligible speech could not be synthesized by decoding acoustic and muscle vibrations of T15's residual speech.

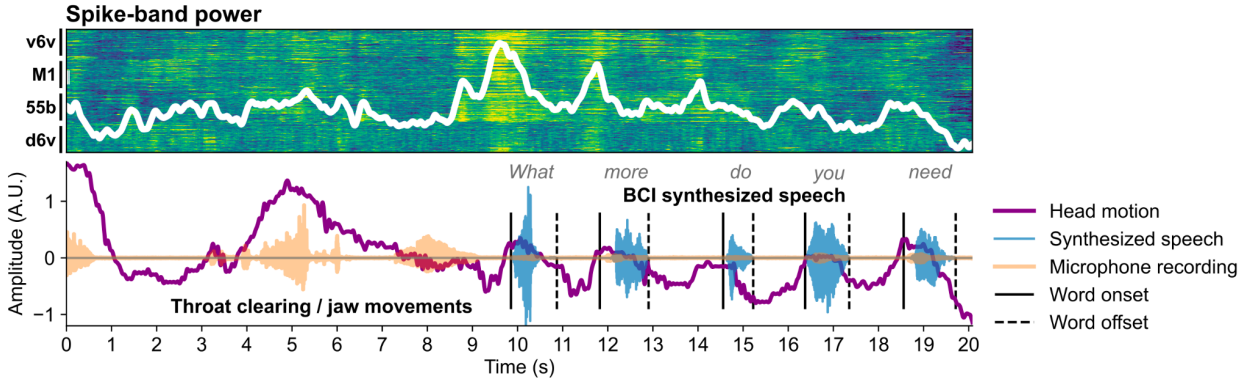
Video: line 1262

Video 15: Intelligible voice cannot be decoded from biosignal recordings of T15's residual speech in contrast to neural activity. This video shows examples of speech synthesized from simultaneous recordings from microphone, stethoscopic microphone, IMU sensor and intracortical neural activity as T15 attempted to speak. Speech synthesized from microphone, stethoscope and IMU biosignals was not intelligible (word error rate for stethoscope decoding: 100%), whereas the voice synthesized from neural activity was more intelligible (word error rate: 43.60%). From post-implant day 482.
Link to view online: https://ucdavis.box.com/s/tz4odg4uefo7ifsilxm92odv1lm5sj03

2.5. Head motion during attempted speech

Next, we tracked T15's head motion during speech to examine whether it could have introduced motion artifacts in the neural signal. We estimated T15's head motion by tracking the movement of one of the NeuroPort pedestals (which are rigidly attached to his skull) in videos recorded during research blocks as shown in **Extended Fig. 11**. We chose to track the pedestal because it is visually distinctive and easier for machine vision algorithms to reliably identify in each frame. We co-registered this movement with simultaneously recorded neural signals and T15's

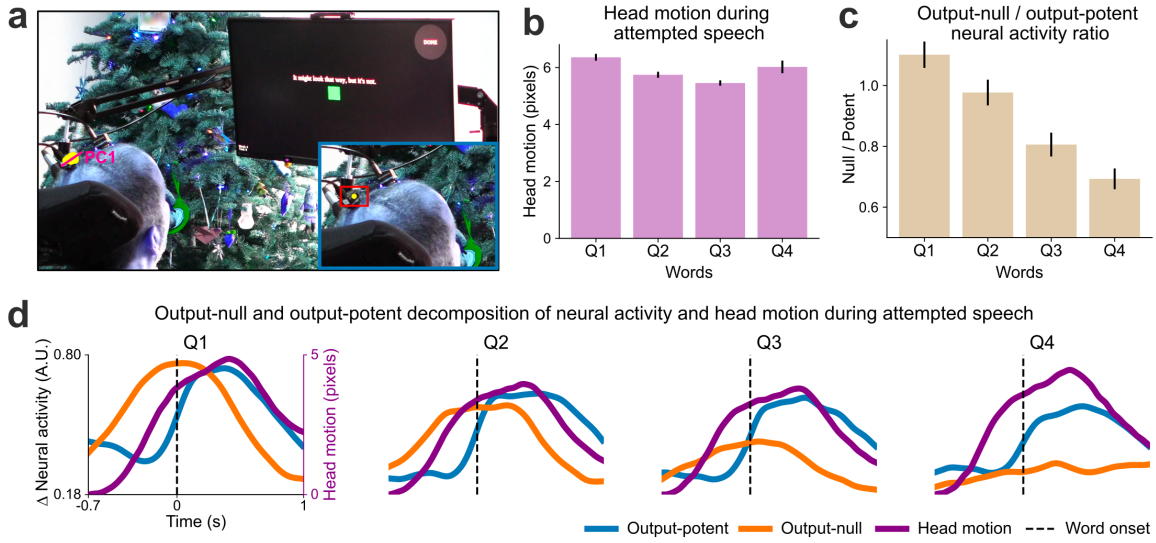
attempted speech audio. T15 made small head motion (~6 pixels in the video captured from approximately 1 meter away) during attempting to utter each word, consistent with retaining very limited neck movement. The amplitude of head motion remained consistent during consecutive words throughout the sentence (**Extended Fig. 11b**). This is in contrast to the neural activity, which decayed over the course of the sentence (with higher neural activity for the words at the beginning of the sentence and lower neural activity for the words at the towards of the sentence; this is characterized by output-null and output-potent decomposition of neural activity during a trial as explained in the main **Fig. 4** and **Extended Fig. 11c**). Different time courses and amplitudes were observed for the output-null and output-potent components of the neural activity versus T15's head motion in different quartiles of the sentence relative to the word onset (**Extended Fig. 11d**). This confirms that the neural activity has different dynamics than the head motion and argues against head motion causing artifacts in the neural activity. Moreover, T15's head motions when he was not speaking did not cause modulation in neural activity (see **Response to Reviewers Fig. 1**), which also suggests that the neural activity was not contaminated by the body movement artifacts.



Response to Reviewers Fig. 1. An example trial showing *(top)* neural activity (white trace shows mean spike-band power across all electrodes), *(bottom)* speech synthesized from neural activity (blue), corresponding microphone recording of T15's attempted speech (orange), and T15's head motion (purple) estimated from simultaneously recorded video. The first 10 s of the trial shows throat clearing and jaw movement accompanied by non-speech vocalization captured by the microphone and corresponding larger head motion. No speech was synthesized during this period when T15 was not attempting to speak. There is also no increase in neural activity during this period and speech-related pattern that can be observed prior to and during utterance of each word.

We have included a new **Extended Data Fig. 11** in the revised manuscript and described its generation in Methods **line 985** as follows:

Extended Data Fig. 11 line 1159:



Extended Data Fig. 11. Head motion during speech and its relationship with the neural dynamics. a. Head motion was tracked from videos by mapping the *x* and *y* positions of the NeuroPort pedestal in each frame (yellow points) using OpenCV. Overall head motion was summarized by the first principal component (pink axis) of *x*-*y* motion. **b.** Small head motion was observed during uttering each word. Head motion remained consistent throughout the attempted speech sentence as measured by the motion from baseline during utterance of words in each of the four word-quartiles, regardless of the length of the sentence. **c.** The ratio of output-null and output-potent components of simultaneously recorded neural activity decayed over the course of the sentence, in contrast to the head motion in (**b**). **d.** Time course and amplitude of the output-null and output-potent components of the neural activity and simultaneous head motion in different quartiles of the sentence. The head motion (purple) follows the output-null activity (orange) but precedes the output-potent activity (blue). The output-null activity decayed over the course of the sentence, whereas the head motion during each word in a sentence remained constant. Taken together, this shows that the neural dynamics do not closely match the head motion time course.

Methods line 985:

We verified that the decaying output-null activity was not merely tracking gross head motion during attempted speech, which could indicate a neural correlate of head movement. To do so, we tracked T15's head motion from videos simultaneously recorded with neural signals and compared changes in head movements with output-null and output-potent neural dynamics over time (**Extended Data Fig. 11**).

Next, we address additional specific questions and feedback from each reviewer:

Referee #1 (Remarks to the Author):

Wairagkar et al present a real-time streaming speech synthesis system in a patient suffering from ALS. The group implanted 4 Utah arrays into the ventral motor cortex of the patient. They feed 600 ms of spike-band power and threshold crossing features into a Transformer Encoder to predict speech parameters that are subsequently transformed into an audio waveform by LPCNet. This whole process is very fast (10 ms delay) and allows for immediate voice feedback for the dysarthric patient. The reconstructed audio is of very high quality and also allows for modulation of pitch and accentuation, which gives the patient access to paralinguistic aspects of speech.

The presented pipeline is a well-designed prosthesis that presents a clear advantage for this patient. In particular, the generation of target speech, which the patient can't produce anymore, is clever. However, I have major concerns whether this prosthesis really relies on neural data:

We thank the reviewer for their careful and thorough assessment of the manuscript. We appreciate the kind words about the low latency [a great deal of engineering effort went into achieving this], high quality of the reconstruction, and the flexibility of pitch and accentuation modulation. We appreciate the importance of rigorously demonstrating that the neuroprosthesis really relies on *neural* data and have conducted multiple new analyses and experiments to demonstrate this, as described in point #2 in our **common responses** above ("*2. Absence of motion artifacts in intracortical neural recordings*").

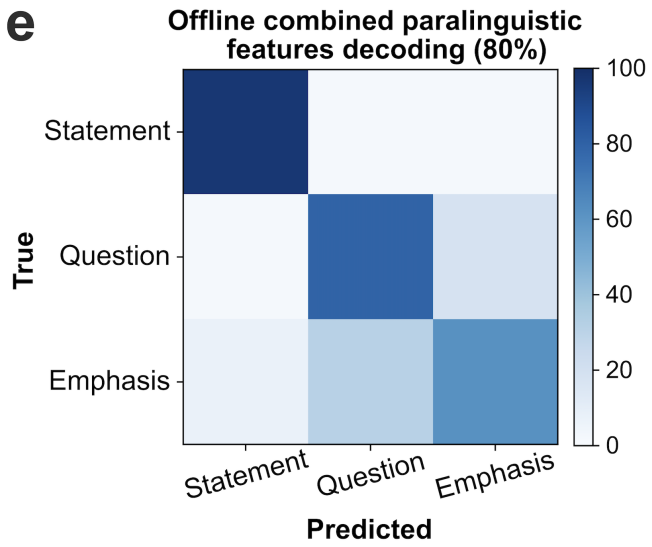
1. The authors note that all four arrays show strongly increased activity for accentuation of words. The authors state: "A functional neuroanatomy result observed in this study that would not be predicted from prior ECoG and microstimulation studies is that the neural activity is correlated with paralinguistic features across all four microelectrode arrays, from ventral-most precentral gyrus to the middle precentral gyrus." (ll.377). To me, this pattern of activity is deeply concerning. In my understanding, it is highly unlikely that all electrodes on all four arrays in the motor cortex should show such a simultaneous strong increase in activity just because a word is accentuated (Fig 3C). The increase is almost two-fold for some arrays (Extended Data Fig 7). Particularly the fact that this finding has not yet been observed in microstimulation studies should make the authors cautious. In my opinion, a far more likely explanation for this increase is contamination through artifacts induced by the face and body movements of the participant. When watching video 11 carefully, it can be seen that the patient moves a lot more during the accentuated words. Such artifacts could easily explain the notable increase of activity on all four arrays. Such speech induced artifacts have been observed before [1-2]. Vibrations during mimed speech production (which is also a condition here) have even been used to reconstruct speech from a vibration sensor to a similarly high degree [3] as reported here. The vibration data collected in [3] also lies in the same frequency range as the extracted spike band power this work relies on.

To address reviewer's concerns about potential artifact contamination in the neural data, we have conducted several analyses and conducted control experiments as described in detail in the **common responses #2 pages 4-14** above. To recap: we performed an acoustic contamination analysis on the neural signals, tested decoders to decode speech from T15's residual movements and acoustics, tracked T15's motion during speech, demonstrated that the BCI does not output voice during non-speech vocalization events such as coughing and jaw

movements, and also showed different tuning on different electrodes to speech. Results from all our analyses consistently and robustly rule out the possibility of artifacts contaminating the neural signals; they thereby confirm that this neuroprosthesis synthesized voice by decoding neural correlates of speech.

We next turn to the reviewer's question about whether it is surprising that all four arrays show modulation to speaking and (mostly) increases to accentuated words. We would argue that this result, while novel (because it's a new task for the intracortical BCI field) are in line with the field's existing understanding of motor cortical function tuning properties at spiking resolution:

- Whilst the homuncular organization across the precentral gyrus has been mostly observed on mesoscale recordings (i.e., on the clinically-relevant scale of millimeters to centimeters), strong evidence in the current literature shows that there is a mosaic of whole body representation intermixed in the cortex. This is particularly evident with intracortical recordings of single units [*Deo et al.* 2024, ref 1*] where the representation of different orofacial regions and limbs was intermixed throughout the precentral gyrus. Our results show that all the four arrays showed speech-related modulation to different extents, which is consistent with broad encoding of the whole body in any given area of motor cortex.
- Additionally, we also observed a rich variety in neural dynamics of all electrodes from different arrays that show complex patterns of tuning to speech (see **common response 2.2 page 6**). So it's not the case that the local neural activity near every electrode is doing the exact same thing (which we agree would be a warning flag that maybe there was a simpler non-neural explanation).
- Prior studies in macaques [*Kaufman et al. 2016*, ref 2*] and our own work in human dorsal precentral gyrus [*Stavisky et al. 2019*, ref 3*] found that the largest common pattern across individual neurons during *any* movement is a condition-invariant signal which is (mostly, but not completely) characterized by firing rate increases. Our results, which show that *on average across electrodes* firing rates go up during speaking (and even more during emphasized words) is consistent with these prior findings.
- In this light, it can be expected to see the contribution from different cortical regions towards changes in intonation during speech that engages several articulatory muscles, larynx and diaphragm forming a complex activity pattern that we were able to decode to identify word emphasis, question intonation or pitch modulation. **Extended Data Fig. 7b** in the manuscript shows different aggregate activity during intonation/pitch modulation between the arrays: for example, 55b array shows delayed increase in the activity as compared to the other arrays. This variety in electrode dynamics rules out the possibility of movement artifacts. It should be noted that the neural activity increase shown in **Fig. 3c** is observed after aligning all trials using dynamic time warping and averaging over all trials which amplifies the effect, yet it still shows different timings for this increase in the neural activity. This is more evidence against artifact contamination in our neural recordings.
- Furthermore, specifically with regard to accentuated words, we have now performed an offline analysis (added in the new **Extended Data Fig. 7** panel **e**) which shows that we can classify between different types of intonation (statement vs question vs emphasis) with 80% accuracy. This classification was achieved with relatively little data; with further optimization and more data we predict we could achieve higher decoding accuracy. This further indicates that there is specific neural information about accentuated words, rather than a general (and perhaps related to overall body movement or artifact) signal:



(**line 1101**) **Extended Data Fig. 7. Encoding of paralinguistic features in neural activity**…
e. Confusion matrix showing offline accuracies for decoding question intonation and word emphasis paralinguistic features together using a single combined 3-class classifier.

[1*] Deo et al., "A mosaic of whole-body representations in human motor cortex", bioRxiv preprint, 2024. DOI: 10.1101/2024.09.14.613041

[2*] Kaufman MT, Seely JS, Sussillo D, Ryu SI, Shenoy KV, Churchland MM. The Largest Response Component in the Motor Cortex Reflects Movement Timing but Not Movement Type. eNeuro. 2016 Sep 1;3(4):1171–97. DOI: 10.1523/ENEURO.0085-16.2016

[3*] Stavisky SD, Willett FR, Wilson GH, Murphy BA, Rezaii P, Avansino DT, et al. Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis. eLife. 2019 Dec 10;8(e46015). https://elifesciences.org/articles/46015

2. To ensure that this prosthesis is not based on artifacts resulting from the patient's movements, I suggest attaching a vibration sensor (stethoscope or IMU) behind the ear of the patient and then recording during attempted speech production. This data can then be likened to the recordings from the Utah arrays and used to predict the LPC coefficents independently. Without this control, it cannot be ruled out that the speech reconstruction in this study is based on artifacts.
On the other hand, this might actually be good news for patients. The results presented in this study and the NEJM study by the same group could benefit patients without the need for brain surgery.

Thank you for your excellent suggestion to conduct control experiments with additional sensors. We conducted new experiments to record multiple simultaneous data streams by attaching a stethoscopic mic on T15's left mastoid to collect vibration data similar to [3] and an IMU sensor with 3-axial accelerometer and gyroscope on his right mastoid and a microphone to record T15's vocalizations along with the neural signals as T15 attempted to speak. Then we trained separate decoders to predict LPCNet coefficients from these modalities independently. Please see the detailed description in the **common response 2.4 page 10** in this document and in the **Extended Data Fig. 6b-d** in the revised manuscript. The signals recorded from these acoustic and vibration sensors could not synthesize intelligible voice (100% word error rate based on

human open transcription) in contrast to neural signals which could produce mostly intelligible voice (43.6% word error rate), which we also demonstrate in a new **Video 15** (https://ucdavis.box.com/s/tz4odg4uefo7ifsilxm92odv1lm5sj03). Thus, we confirmed that the voice neuroprosthesis was indeed based on neural activity and did not decode artifacts from T15's residual speech or movements.

As an aside, we also wish it had been the case that we could restore accurate communication using just non-invasive signals. That would indeed hold remarkable benefit for patients and such a therapy could be scaled much faster and more affordably. Unfortunately, our data indicate that we do need to perform brain surgery to implant electrodes to get this level of speech restoration. Thus, we will keep pushing on that to maximize the benefits of this technology.

Further points:

3. * Given the very high quality of the output, the intelligibility assessment is not well suited. Clearly, the six-choice sentence selection is far too easy, as the authors observed ceiling effects of almost 100% accuracy. I think an open transcription task on the reconstruction would provide a better assessment.

We agree with the reviewer that open transcription of the BCI-synthesized voice by naïve human listeners is a better evaluation metric. We have conducted open transcription of brain-to-voice synthesized speech as detailed in the **common response #1** on **page 1** of this document and also added this to the main results (**Fig. 2l**) in the manuscript. Open transcription yielded phoneme error rate of 34.00% and word error rate of 43.75% for synthesized voice. To our knowledge, this is the lowest word error rate achieved by a speech synthesis BCI in the literature. For comparison, T15's residual speech had a word error rate of 96.43%. This shows that the BCI voice is substantially more intelligible than T15's dysarthric speech.

4. * A speaker identification task should be conducted (using listeners on Mechanical Turk, again) to assess whether the reconstructed voice (in the own-voice condition) really matches the participants old voice.

We have included a new **Video 16** (https://ucdavis.box.com/s/8zbkvtbau04543rxiwjan8i4k40eo6xy) in the manuscript that compares T15's pre-ALS voice that was used to train the voice cloning model with the target cloned voice used for BCI training with the BCI-synthesized own-voice to enable readers to self-assess the similarity between these voices.

We obtained T15's subjective feedback by asking what he thought of the own-voice BCI and his response was "*it made me feel happy and it felt like my real voice*" (included in the manuscript **line 223**, **Fig. 2e** and shown in **Video 6**). Additionally, to the question whether the BCI voice sounded like his own voice, he responded "*yes it did, as much as it could at this point*". Note that T15 responded to these questions using the brain-to-voice BCI and he confirmed his responses through text decoding.

We have included this new video in the manuscript in **line 1270** and has been cited in Methods **line 1270.**

Video 16: line 1270

> **Video 16: Comparison of T15's pre-ALS voice with the "own-voice" synthesis by brain-to-voice BCI.** This video shows examples of (1) T15's pre-ALS voice; (2) voice cloned by the StyleTTS 2 model which was trained on his pre-ALS voice; (3) target audio generated using this cloned-voice, time-aligned with neural signals during attempted speech used as training data for personalized "own-voice" BCI speech synthesis decoder; and (4) "own-voice" synthesized by the personalized brain-to-voice decoder in closed-loop from neural activity.
> Link to view online: https://ucdavis.box.com/s/8zbkvtbau04543rxiwjan8i4k40eo6xy

Methods: line 1270

> **Video 16** shows the details of voice cloning pipeline and comparison of T15's pre-ALS voice with own-voice BCI synthesis.

5. * I suggest to additionally cite the first paper that used a Transformer-Encoder architecture to reconstruct spectrograms from neural data [4] as the proposed pipeline is very similar.

Thank you for suggesting this additional reference that shows other studies using a Transformer-Encoder architecture for decoding neural (in that case, ECoG) signals into speech spectral features. We now cite it in the Introduction. While we agree that this contemporaneous work is relevant, our approach was not inspired by the Shigemi et al. study; rather, we had already started prototyping this architecture with previously-recorded speech-related neural activity from hand motor cortex, as described in our conference paper (Wairagkar 2023) that was published online in the same month as Shigemi et al. 2023.

The updated snippet from the Introduction **line 54** section reads:

> A growing literature of studies have reconstructed voice offline from able-bodied speakers using previously recorded neural signals measured with electrocorticography (ECoG)[5-12], stereoencephalography (sEEG)[13], and intracortical microelectrode arrays[14,15].

(ref 12 is Shigemi et al. May 2023, Ref 15 is Wairagkar M, Hochberg LR, Brandman DM, Stavisky SD. Synthesizing Speech by Decoding Intracortical Neural Activity from Dorsal Motor Cortex. In: 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER). May 2023. p. 1–4.)

6. * To put the MCD results of the acausal pipeline (l. 148) in perspective for the reader, it would be good to report the results of the causal one here, too.

Thank you for pointing out that it would help the reader to compare the causal and acausal accuracy by also reporting the causal MCD value. The updated Results **line 166** sentence now says:

> As expected, acausal synthesis – which benefits from integrating over the entire utterance – generated high quality voice (Mel-cepstral distortion: 2.4±0.03 versus 2.9±0.5 for causal synthesis); this result illustrates that instantaneous voice synthesis is a substantially more challenging problem.

[1] Bush, A., Chrabaszcz, A., Peterson, V., Saravanan, V., Dastolfo-Hromack, C., Lipski, W. J., & Richardson, R. M. (2022). Differentiation of speech-induced artifacts from physiological high gamma activity in intracranial recordings. NeuroImage, 250, 118962.
[2] Roussel, P., Le Godais, G., Bocquelet, F., Palma, M., Hongjie, J., Zhang, S., ... & Yvert, B. (2020). Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. Journal of Neural Engineering, 17(5), 056028.
[3] Shah, N., Sahipjohn, N., Tambrahalli, V., Subramanian, R., & Gandhi, V. (2024). StethoSpeech: Speech Generation Through a Clinical Stethoscope Attached to the Skin. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(3), 1-21.
[4] Shigemi, K., Komeiji, S., Mitsuhashi, T., Iimura, Y., Suzuki, H., Sugano, H., ... & Tanaka, T. (2023, June). Synthesizing speech from ecog with a combination of transformer-based encoder and neural vocoder. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Referee #2 (Remarks to the Author):

Summary of the key results
The manuscript by Wairagkar ea reports on realtime decoding of attempted or mimed speech by a person with severe paralysis due to ALS. Small chunks of Utah-array data (256 electrodes) were directly converted to audio with a delay of some 50 ms, and the result was promising in terms of intelligibility. Moreover, authors could identify changes in emphasis/intonation and pitch, which likely further enriches the users' experience of speaking via a computer.

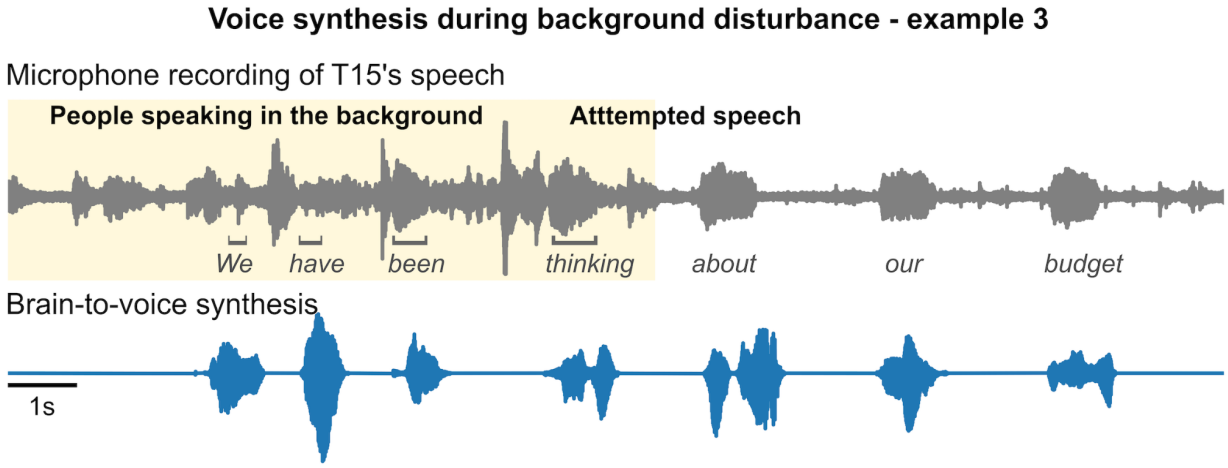
Originality and significance: if not novel, please include reference
1. The results are very important in that the user can express him/herself vocally, at a speed that would allow participation in conversation with company. At this point it is not clear yet whether decoding will be feasible outside of research context, for instance with other peoples voices potentially degrading decodability of own speech. Moreover, it is not clear whether decoding is equally feasible in people who can not mime due to facial paralysis, such as in the case of late-stage ALS. Nevertheless, the results pose a significant advance in Brain-Computer Interface research in terms of speed and intelligibility. I do have some concerns, notably about the measure of intelligibility, amount of training required plus stability of decoders over time, correlated brain signal features for paralinguistic variables and generalizability of findings to other people with severe paralysis.

We thank the reviewer for their careful and thorough assessment of the manuscript. We appreciate their positive description of the work, including the intelligibility, the importance of technology that enables people to express themselves vocally, and that the ability to change emphasis/intonation and pitch likely enriches the user's experience. We also appreciate the reviewer raising questions and concerns and believe we have been able to address these (as described in the point-by-point responses below) in a way that has improved the manuscript.

With respect to whether decoding is feasible at the same time as other people are speaking, we draw the reviewer's attention to this example in the new **Extended Data Fig. 5a** which shows that voice synthesis is not falsely triggered by hearing the start of people speaking in the

background and then works while the participant is speaking at the same time as background conversation:

Extended Data Fig. 5a:



The reviewer's comment also made us realize that we had overlooked mentioning an important element of this study: all of these data were collected in the home of the participant (i.e., in the "real world" and not in the lab). We have added the following text to the Methods **line 615** to explain this:

> This study comprises multiple closed-loop speech tasks used to develop and evaluate a voice synthesis neuroprosthesis. Research sessions were conducted in the home of the participant. They were structured as a series of blocks of ~50 trials of a specific task.

Anecdotally, we have seen the speech neuroprosthesis work despite 1) people talking in the background, 2) a blaring fire alarm [we did check that there was no actual fire before resuming research], 3) a parade of motorcycles outside the front window, and 4) blaring music from cars stopped outside [it is a sometimes-loud urban environment].

2. Data & methodology: validity of approach, quality of data, quality of presentation
The analytical approach is appropriate and straightforward, although I have concerns about how intelligibility was quantified. Data quality and presentation is quite adequate

Thank you for the positive feedback. Regarding quantifying intelligibility, we have now evaluated synthesized speech through an open transcription task by naïve human listeners as detailed below and in the **common response #1** above on **page 1** of this document.

3. Appropriate use of statistics and treatment of uncertainties
This is a single-subject study hence statistics are only within-subject. They are appropriate in terms of statistical evaluations of reported results of experimental tests

Thank you.

4. Conclusions: robustness, validity, reliability
Results seem robust, but there are some unclarified issues regarding training data, stability of decoders and generalizability to people with facial paralysis

We have now addressed these concerns in detail below.

5. Suggested improvements: experiments, data for possible revision
The measure for intelligibility is not quite adequate, it is suggested to apply a more rigid and quantifiable measure for each single word

We have now evaluated synthesized speech through an open transcription task by naïve human listeners as detailed below and in the **common response #1** above on **page 1.**

6. References: appropriate credit to previous work?
References are adequate

We are relieved to hear that we did not overlook essential previous work.

7. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions
The manuscript is written concisely and clearly. Conclusions may need additional information

We thank the reviewer for this useful feedback. We have addressed all comments below and made the required changes in the manuscript, including providing additional information in the conclusions.

Major comments

8. Regarding a key measure of performance, intelligibility, I am rather surprised at the approach the authors took. Intelligibility was assessed by identification by 15 volunteers ('We quantified intelligibility by asking 15 human listeners to match each of the 956 evaluation sentences with the correct transcript '(choosing from 6 possible sentences of the same length)'. This is a rather forgiving test since with only a few intelligible words (or even one if sentences differ on the first) a whole sentence can be identified out of 6. Yet authors then state that 'These results demonstrate that the real-time synthesized speech ... can be identified by non-expert listeners.' This seems a rather big leap in interpreting results. To get a better estimate of intelligibility I would prefer to see a test where volunteers do not choose between sentences but write down for every single word what they think they heard. Tis can then be compared to the prompted word and provides a much more informative measure. A need for a better measure is further indicated by the conclusion in Discussion that 'The resulting voice was often intelligible' which contrasts with the near 100% performance on the 6-choice test.

We agree with the reviewer that multiple-choice identification is not an ideal metric, especially given that (as the reviewer noted) we were hitting ceiling effects. As per the reviewer's suggestion, we have now conducted open transcription tests of brain-to-voice synthesized speech as detailed in the **common response #1** on **page 1** of this document and added these results to the main results (**Fig. 2i**) in the revised manuscript. Open transcription yielded

phoneme error rate of 34.00% and word error rate of 43.75% for synthesized voice. To our knowledge, this is the lowest word error rate achieved by a speech synthesis BCI in the literature. For comparison, T15's residual speech had a word error rate of 96.43% as transcribed by the same listeners. This shows that the BCI voice is substantially more intelligible than T15's dysarthric speech.

9. Authors do not mention the fact that T15 used movement of his articulators to generate the brain signals for brain-to-speech decoding in the Discussion section. As I assume authors are aware of, somatosensory feedback signals are present in primary motor cortex (eg doi.org/10.1007/BF00237846, doi:10.1152/jn.00949.2014, doi.org/10.1101/2023.08.05.552025) and may well contribute to decoding of speech. Authors should discuss this and add to the Limitations section as a constraint to generalizability the potential ramifications for decodability of speech in people with facial paralysis or complete paralysis (eg LIS)

T15 retained limited orofacial movement, but not enough control of articulators to produce intelligible speech (naïve human listeners found his speech unintelligible and transcribed it with a word error rate of 96.43%). However, as pointed out by the reviewer, it is currently unknown to what extent these residual articulatory movements (including their somatosensory neural correlates) contribute to brain-to-voice synthesis and how well this method can generalize to people with complete paralysis. We have now highlighted this limitation of the study in the discussion section where we had previously included the limitation of generalizability to different speech loss etiologies as follows in **line 421**:

> This study included a single participant with ALS who retained limited articulatory movement and the ability to vocalize (unintelligibly), with accompanying sensory feedback. It remains to be seen whether similar brain-to-voice performance will be replicated in additional participants, including those with other etiologies of speech loss or people in late-stage ALS with complete paralysis and locked-in state. The participant's ALS should also be considered when interpreting the study's scientific results. Encouragingly, however, prior studies have found that neural coding observations related to hand movements have generalized across people with ALS and able-bodied animal models[34] and across a variety of etiologies of BCI clinical trial participants[35,36]. Furthermore, the phonemic and paralinguistic tuning reported here at action potential resolution has parallels in meso-scale ECoG measurements over sensorimotor cortex in able speakers being treated for epilepsy[24,29].

10. A frequent occurring challenge with speech decoding is training and calibration, an issue that is typically addressed in BCI publications. I am sure it is distributed across the methods section but can the authors clearly indicate the information readers need to assess how much time the participant spent generating data for model training, and stability of decoders over time. Include a timeline with days starting on the first day of generating training data and ending with the day of the last test, and make clear the time (days) between last decoder model training and the test.

We thank the reviewer for these suggestions to make the methodological details of this study more clear and useful. We have now included an **Extended Data Table 1** (**line 1173**) in the manuscript summarizing the details of the training data used for training the decoders shown in

the core brain-to-voice results in **Fig. 2**. This table includes the time duration between decoder training and evaluation day as follows:

Extended Data Table 1: Data used for training the brain-to-voice decoder

Evaluation day (post-implant)	Most recent decoder training day	Days between decoder training and evaluation	Cumulative number of training sentences	Cumulative hours of training data
25 (1st session)	25	0	180	0.5
109	104	5	4,140	17.4
110	109	1	4,389	18.7
125	123	2	4,648	20.1
139	132	5	4,936	21.6
165	153	12	5,921	26.9
172	165	7	6,254	28.5
174	172	2	6,552	30.1
179	174	5	6,778	31.3
181	179	2	7,077	32.8
186	181	5	7,279	34.0
188	186	2	7,589	35.3
193	188	5	7,890	36.4
195	193	2	8,315	37.8

Details of decoder training were included in a separate Methods section starting from **line 709.**

On the first session of neural recording, we collected 180 open-loop trials of attempted speech from a 50-word vocabulary to train the decoder and were able to synthesize voice in closed-loop with audio feedback on the same day. Although the closed-loop synthesis on this data was less intelligible due to the model not being optimized on the first day, we later demonstrated offline that with an optimized model, we could achieve intelligible synthesis with this small amount of neural data and a limited vocabulary (**Video 14**).

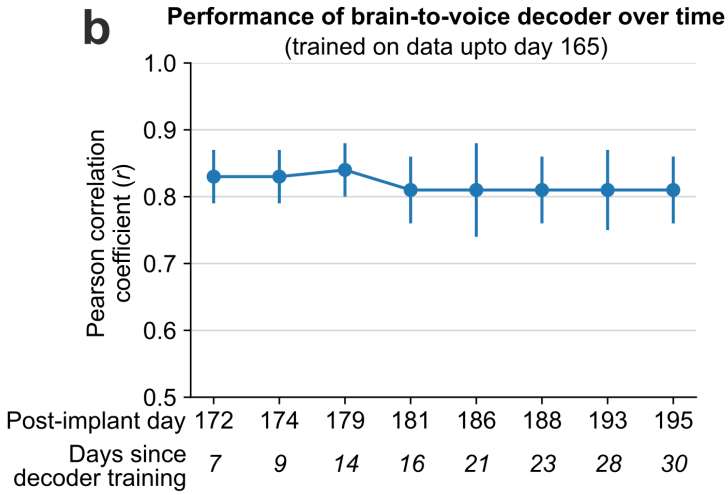
In subsequent sessions, we collected more attempted speech trials with a large vocabulary and iteratively optimized our brain-to-voice decoder architecture to improve the synthesis quality. Here, we report the performance of closed-loop voice synthesis from neural activity using the “final” brain-to-voice decoder architecture for predetermined evaluation sessions. For each of these sessions, the decoder was trained on all the data collected up to one week prior (total of ~4,100-8,300 trials). Further details of the data

used for training the brain-to-voice decoders in each of the sessions shown in **Fig. 2** are provided in **Extended Data Table 1**.

We evaluated the longevity of decoder over time offline by fixing a pre-trained brain-to-voice decoder (trained on post-implant day 165) and assessed its performance on data collected in the sessions over a period of one month (shown in the figure below). We can observe that there is a dip in the performance of the decoder (measured by Pearson correlation coefficient across 40-Mel frequencies by excluding silences between the words) after 15 days. We can also subjectively corroborate this observation that there is a noticeable degradation in performance after 15 days.

We have included this information in the manuscript in **Extended Data Fig. 3b** and Results **line 162.**

Extended Data Fig 3b: line 1031



Extended Data Fig. 3. Additional BCI speech synthesis performance metrics…
b. Performance of brain-to-voice decoder measured over time by evaluating neural trials from different sessions offline with a fixed decoder. Decoder trained on post-implant day 165 was fixed and used to synthesize voice offline from neural trials collected in sessions over the next month. Performance was measured by computing Pearson correlation coefficient between the target speech and the synthesized speech across 40 Mel-frequencies after removing silences between words. A noticeable decline in brain-to-voice performance was observed after approximately 15 days.

Results: line 162

The performance of the brain-to-voice decoder over time is shown in **Extended Data Fig. 3b**.

11. Fig3: loudness, intonation and pitch were studied separately, and it seems they are all associated with amplitude of spike band power. Could they measure one and the same phenomenon (ie a single paralinguistic feature)? Can be decoded independently if co-occurring in attempted/mimed speech? please clarify

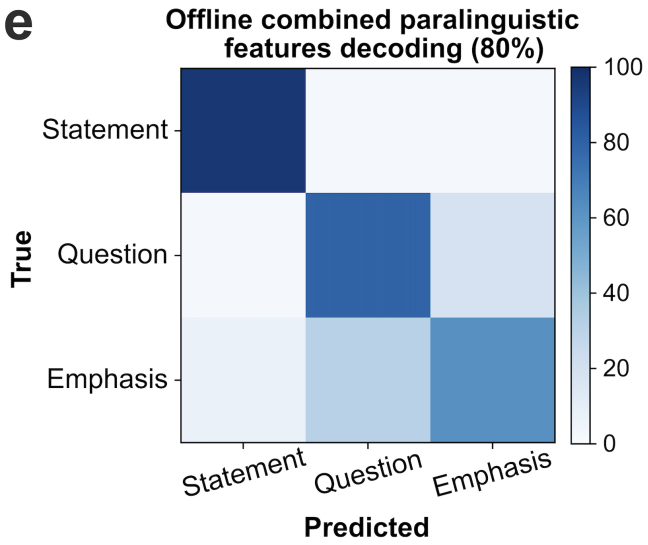
The reviewer raises the interesting question about whether (and to what extent) intonation,

loudness and pitch can be disentangled in neural signals. Indeed, in our current tasks, both the emphasized word or question-asking likely involve a combination of changing loudness and pitch.

As a step towards this which we could do with our existing data, we have now performed a new offline analysis where we classify word emphasis versus asking a question versus a statement from neural data (these behaviors were performed in the same research session). This is now mentioned in the revised Results **line 294**:

> As a proof-of-principle that these paralinguistic features could be captured by a speech neuroprosthesis, we trained two separate binary decoders to identify the change in intonation during these question intonation and word emphasis tasks. We then applied these intonation decoders in parallel to the brain-to-voice decoder to modulate the pitch and amplitude of the synthesized voice in closed loop, enabling T15 to ask a question or emphasize a word (**Extended Data Fig. 8**). **Fig. 3d** shows two example closed-loop voice synthesis trials, including their pitch contours, where T15 spoke a sentence as a statement and as a question. The synthesized speech pitch increased at the end of the sentence during question intonation (**Video 10**). **Fig. 3f** shows two example synthesized trials of the same sentence where different words were emphasized (**Video 11**). Across all closed-loop evaluation trials, we decoded and modulated question intonation with 90.5% accuracy (**Fig. 3e**) and word emphasis with 95.7% accuracy (**Fig. 3g**). The neural correlates of emphasized words, question words, and statement words were distinguishable from each other, with 80.0% offline classification accuracy using a 3-class decoder (**Extended Data Fig. 7e**).

Here is the new **Extended Data Figure 7e** and caption (**line 1101**):



Extended Data Fig. 7. Encoding of paralinguistic features in neural activity

…

e. Confusion matrix showing offline accuracies for decoding question intonation and word emphasis paralinguistic features together using a single combined 3-class classifier.

The updated Methods **line 841** describe the analysis:

> In an offline analysis, we tested whether a single combined paralinguistics decoder could classify between no change in intonation (class 0), question intonation modulation (class 1) and word emphasis (class 2). We trained a three-class logistic regression classifier using the same training data as above. We generated training samples by using the neural spike-band power 600 ms prior to the word onset up to word offset and by sampling it with sliding windows of 600 ms shifted by 10 ms to get the same sized feature vector as above. We applied this combined three-class decoder on individual words (600 ms sliding windows applied from a time range consisting of -400 ms to 400 ms from the onset) from the same evaluation data as closed-loop paralinguistic features modulation. The most frequently predicted class during this time range was selected as the final class for that word.

Given that the manuscript was already quite packed in terms of content, and we have now added new analyses and datasets in this revision, we respectfully request that a deeper investigation into this question be left for future work. We believe that the current results are important as they 1) establish that paralinguistic features can be decoded from intracortical arrays in these four brain locations, and that 2) these signals can be decoded accurately and fast enough to be used in a low-latency closed-loop speech neuroprosthesis. We hope that in future work we (or other groups that can build on this study) can dig deeper into teasing these voice features apart. We believe that the present study will help motivate and facilitate such work.

Minor comments

12. Line 52: imagining instead of imaging

Thank you for pointing out this typo. We have now corrected it in **line 57.**

> … synthesize unintelligible speech during miming, whispering or imagining speaking tasks online[8,16]…

13. General: please elaborate on what 'causal' and 'acausal' decoding mean other than (near) realtime. Perhaps those terms don't add anything to the discourse. Surely brain signals constitute cause to speech decoding even if not immediate?

Thank you for pointing out that we should briefly unpack words that may be jargon in a broad-audience journal like Nature. 'Causal' and 'Acausal' are widely used in the fields of electrical engineering, BCI, machine learning and related fields. Here, 'causal' means that we are not using neural data ahead of the point in time that the user is trying to speak. By contrast, 'acausal' means that we would have access to future data for the decoder. Since we are only decoding data up to the point of time when the user is actually trying to speak, we are performing causal decoding. By contrast, another approach to perform speech decoding (like past methods in the speech BCI literature) would have been to collect several seconds of data after the user had completed an utterance (to “look ahead”), and use the entire time epoch to perform acausal decoding. We have added parenthetical explanations in the revised manuscript’s Introduction **lines 46, 61** and **65** upon the first use of each of these terms:

These additional capabilities can be restored with a “brain-to-voice” BCI that decodes neural activity into sounds in real-time that the user can hear as they attempt to speak. Developing such a speech synthesis BCI poses several unsolved challenges: the lack of ground truth training data, i.e., not knowing how and when a person with speech impairment is trying to speak; causal (only using past neural signals) low-latency decoding …

While the aforementioned studies were done with able speakers, a study [3] with a participant with anarthria adapted a text decoding approach to decode discrete speech units acausally (i.e., using neural data from both before and after a given utterance) at the end of the sentence to synthesize speech from a 1,024-word vocabulary…

In this work, we sought to causally synthesize voice continuously and with low latency from neural activity as the user attempted to speak, which we refer to as “instantaneous” voice synthesis to contrast it with earlier work demonstrating acausal delayed synthesis.

14. Line 65 : ‘To overcome this, we generated synthetic target speech waveforms from the prompt text' Please explain what the prompt text refers to

Thank you for pointing out that “prompt text” is not an intuitive phrase. We meant the cue text prompted on the screen during speech trials. For clarity, we have changed this to ‘**text cued on-screen**’ in updated **line 73**.

To overcome this, we generated synthetic target speech waveforms from the text cued on-screen and

15. Line 180: 'The accuracy of his free response synthesis was slightly lower than that of cued speech'. How was this determined, did the participant also produce or indicate the text he wanted to voice?

Thank you for pointing out that our methods were not sufficiently clearly described here. During spontaneous speech trials, along with brain-to-voice synthesis, we also ran a separate brain-to-text decoder (described in Card et al. 2024, *NEJM, https://www.nejm.org/doi/full/10.1056/NEJMoa2314132*) which displayed decoded words on the screen as T15 attempted to speak. This text output (almost like closed-captioning) subsequently allowed T15 to confirm what he said or – if there were mistaken words – correct the text decoding. This provided us with ground truth intended text for each spontaneous speech trial. We used this ground truth text to generate target speech for spontaneous free response trials which allowed us to evaluate the performance of synthesized speech. We have now clarified this as follows in the revised methods:

Results line 186:

We used a simultaneous text decoder[1] and confirmed all responses with T15 to get the ground truth text of what he said, which was used to generate target speech and evaluate synthesized speech.

16. Fig 3: previous studies suggest that speed of speaking affects neural activity (eg doi: 10.1016/j.neuron.2016.01.032, www.nature.com/articles/s41583-020-0304-4), did the authors assess decoding performance differences for slow and fast speech? This could be important for a users' performance optimisation.

We conducted a task where T15 was asked to speak sentences in slow speed or fast speed and the brain-to-voice decoder was able to synthesize voice reliably at both speeds at the same pace as T15's attempted speech (results are shown in **Fig. 3a,b**). For this task, we let T15 self-determine the fast or slow speeds at which he could speak. During slow speed trials, it took him on average 1.46±0.31 s to utter each word (Pearson correlation 0.68±0.07) and for fast speed trials, it took him 0.97±0.19 s per word (Pearson correlation 0.75±0.08). Due to ALS, in general, T15 was unable to speak at a fast pace and his speech was significantly slower than healthy speech. He spoke at the rate of ~32 words/min during research tasks. Hence, we cannot determine how the decoder will perform for significantly faster speech rates (e.g.,150 words/min) but believe this is an important topic for future work. We have added text to the Limitations to note this open question (**line 437**):

> It remains an open question whether consistent long-term use will result in improved accuracy due to additional training data and/or learning and whether performance will be sustained at faster speaking rates.

If allowed to speculate: we predict this same exact strategy would work well at faster rates, especially when anticipating having higher electrode-count neural interfaces in the near future (which would provide higher signal-to-noise ratios for estimating speech intent). Our brain-to-voice decoder is built in a way that it causally synthesizes voice every 10 ms regardless of the speed of speech, so from a machine learning/software perspective it can definitely "keep up". During all our trials, it followed the exact pace of T15's attempted speech without any degradation in performance due to variability in his speaking speeds.

Referee #3 (Remarks to the Author):

The study reports an instantaneous voice synthesis neuroprosthesis. Previous studies have successfully decoded imagined speech from intracranial neural signals, but the output is generally text. The current study, however, decodes the speech sound and therefore advances previous studies in two aspects: (1) decoding the timing of syllables and (2) decoding pitch information that is not present in the text. I think the two advances are of potential scientifical importance and are critical for potential application of the method in ALS patients. However, I have major concerns about what signals are actually decoded and whether the results are appropriately evaluated.

We thank the reviewer for their careful and thorough assessment of the manuscript. We were glad to read that they feel that our advances in decoding the linguistic (syllable) and paralinguistic content of speech are critical for potential application of the method in ALS patients. As described in more detail below, we have responded to the reviewer's concerns about the nature of the decoded signal and how the synthesis quality was evaluated.

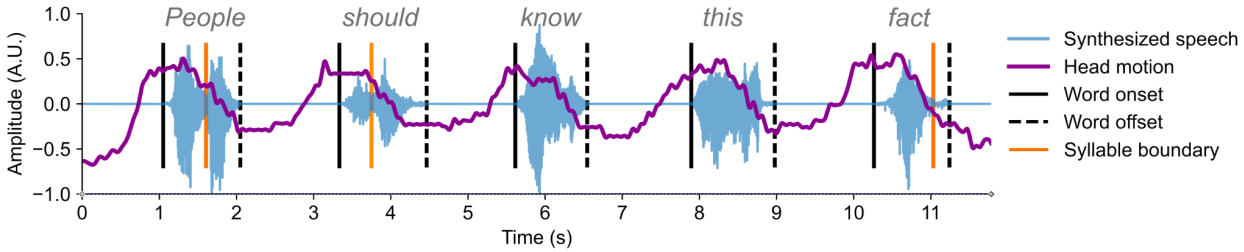
1. Decoding of syllable timing: From the videos, it is clear that the patient moves whenever trying to produce a word. On the positive side, I wonder if this signal could be combined with the

BCI to further improve performance. For example, I wonder if the syllable decoder can be initially trained based on body movements, i.e., treating the onset of each stereotypical body movement as the onset of a word when training the syllable decoder. The authors may also consider to directly decode syllable boundaries based on the video and replace the neural syllable decoder with the video syllable decoder to see which achieves better speech synthesis performance.

On a slightly negative side, I wonder if the body movements are driving the 'null space' activity. In other words, the 'null space' activity is not related to speech per se but is related to the overt body movement, which is absent or much weaker in heathy individuals. Furthermore, I wonder the syllable decoder could actually read out signals controlling such overt body movements, instead of the covert syllable boundaries. It's absolutely fine in terms of the application but scientifically I wonder if the authors could clarify what kind of signal contributes to syllable boundary decoding. I also wonder if some channels behave like the output-potent neural activity.

We segment data into syllable boundaries using neural neural decoder output only (i.e., we use the synthesized speech itself to identify word and syllable boundaries). Decoding these timings using body movement is an interesting idea, however our rationale for solely using neural activity (without relying on residual movement) was to build a decoder that could generalize to people with paralysis who may not be able to move as much (if at all). People with ALS eventually lose their ability to move as the disease progresses. Hence, our goal was to develop a brain-to-voice decoder that can just utilize speech correlates in neural activity.

To assess the extent of T15's movement during attempted speech, as suggested by the reviewer, we tracked his head movement from videos recorded during research blocks (see the details in **common response 2.5** on **page 12** of this document). We noticed that while T15 shows head movement during utterance of each word, his movements are not consistent from trial-to-trial or precise enough to extract timings for syllables reliably. This can be seen in the example trial shown in the **Response to Reviewers Fig. 2** below. The trace in purple shows head motion in this trial; the blue trace is the synthesized speech generated by neural decoder; and vertical lines are precise word and syllable boundaries detected using the synthesized speech (our current approach). As one can see, the head motion precedes the speech by a variable amount of time (not uniformly aligned with speech) and hence it is not sufficient to identify the exact onset and offset of each word. Moreover, it does not contain any finer-grained speech related information (there aren't consistent dips between syllables) so identifying syllable boundaries within a word is not feasible. For training a brain-to-voice decoder that produces low latency speech synthesis (<10 ms), we require very precise timings of syllable-like segments to generate accurate target speech, which we believe is not possible using gross body/head movement (at least in this participant). Moreover, T15 also showed movement during non-speech events such as coughing, throat clearing etc. during attempted speech trials (see example in **common response 2.3** on **page 8** and **Response to Reviewers Fig. 1** on **page 13** of this document).



Response to Reviewers Fig. 2. An example trial showing synthesized speech from neural activity (blue), automatically generated word and syllable segmentation (vertical lines) based on neural decoder output and T15's head movement during this trial as estimated from the simultaneously recorded video (purple).

Our syllable-like segment detection method for target speech generation is detailed in the Methods section **line 654** as follows:

> In subsequent sessions, we relied solely on neural data to estimate syllable boundaries: we used a brain-to-voice model trained on past neural data to synthesize speech and used its envelope to automatically segment syllable boundaries[40], which were manually reviewed and occasionally adjusted if required.

To examine whether the body movements were driving null space activity, we compared the body movement with simultaneously obtained voice output-null and output-potent neural subspace dynamics as detailed in **common response 2.5** on **page 12**. We observed that the amplitude of T15's head movement remained consistent across each word throughout the sentence. On the other hand, output-null activity decayed over the course of the sentence, regardless of the length of the sentence. This can be better appreciated by plotting average head motion over four quartiles of sentences with different length (i.e., the first one-fourth of the words in a given sentence, the second one-fourth, and so on) and also the speech null/potent ratio over same sentences. A clear difference in the dynamics of head motion and null/potent activity can be observed. This suggests that null-space activity cannot be explained by movement. Moreover, speech could be synthesized from null activity independently (main results **line 340**) with a Pearson correlation coefficient of 0.84±0.05; this also confirms that the output-null component contains speech-related information (rather than e.g., just head movement-related information).

Results: line 340

> Both subspaces contained substantial speech-related information: the Pearson correlation coefficients of decoding speech using only the output-potent dimensions (which captured 2.5% of total variance) or only the output-null dimensions (97.5% of total variance) were 0.79±0.08 and 0.84±0.05, respectively.

Details of this analysis are added in the manuscript as described in **common response 2.5** on **page 12**.

2. Speech rate: A sharp difference between the decoded speech and normal speech is the speed. Normal people speak 5-8 syllable per second but the decoded speech is at most 1-2 syllable per second, with long pauses between words. I don't think it's a problem but it needs to be discussed – Why did the patient speak so slowly and lose the ability to speak fluently? Did the patient try to slow down to cooperate with the BCI or which part of the brain might have been affected by the disease so that the person can no longer attempt to speak fluently?

The reviewer raises an excellent question (which we bet many readers will have), and we now recognize that we should provide more context. T15 has severe dysarthria (mixed upper and lower-motor neuron dysarthria and an ALS Functional Rating Scale Revised (ALSFRS-R) score of 23 [range 0 to 48 with higher scores indicating better function) and is unable to speak intelligibly. He is non-ambulatory and retains very limited orofacial movement (as described in the Methods section **line 565**). Because of his severe ALS-induced paralysis, he faces significant difficulty in moving his articulatory muscles while attempting to speak. Hence his pace of speech is significantly slower than healthy speakers. Often, individuals with ALS undergo training with speech and language pathologists to learn to speak slower such that they can retain intelligibility for as long as possible. They also face difficulty with breath control that leads to frequent pauses in speech before uttering each word. These factors have contributed to slowing down T15's speech significantly. Our arrays were placed in the speech motor cortex that supposedly controls the speech articulators, and hence to get good signal-to-noise ratio during motoric speech, we instructed T15 to speak as clearly as possible by enunciating each word fully to the best of his ability. This was effortful for him (example of T15's attempted speech is shown in **Video 1**). Due to all these factors, T15's speech rate was slower.

We conducted a task where we asked T15 to speak sentences at fast speed or slow speed (we did let T15 self-determine how fast or slow he could speak) as shown in **Fig 3a and b**. This showed that during fast speed, he spoke at the rate of 0.97±0.19 s per word and during the slow condition, he spoke at a rate of 1.46±0.31 s per word. The brain-to-voice BCI was able to synthesize speech at both of these speaking rates. To summarize, the limiting factor for his speech rate was ALS. We have highlighted this in Results **line 145** as follows:

> The instantaneous voice synthesis accurately tracked T15's pace of attempted speech (**Extended Data Fig. 4a**), which – due to his ALS – meant slowly speaking one word at a time with pauses in between words. These results demonstrate that the real-time synthesized speech recapitulates the intended speech to a high degree, and can be identified by non-expert listeners.

We have also added this context to the Methods section **line 572** when introducing the participant:

> T15 was a left-handed 45-year-old man. His ALS symptoms began five years before enrolment into this study. At the time of enrolment, he was non-ambulatory, had no functional use of his upper and lower extremities, and was dependent on others for activities of daily living (e.g., moving his wheelchair, dressing, eating, hygiene). T15 had mixed upper and lower-motor neuron dysarthria and an ALS Functional Rating Scale Revised (ALSFRS-R) score of 23 (range 0 to 48 with higher scores indicating better function). He retained some neck and eye movements but had limited orofacial movement. T15 could vocalize but was unable to produce intelligible speech (see **Video 1**). He could be interpreted by expert listeners in his care team, which was his primary mode of communication. Due to ALS, T15's pace of speech was significantly slower than healthy speakers. Furthermore, he had previously undergone training with speech language pathologists who taught him to speak slower in an effort to make his residual speech more intelligible.

3.a Results evaluation: The correlation coefficient between target and synthesized speech is used to evaluate speech decoding, but this measure is not particularly informative. First, the high correlation is driven by the long pauses in patient speech. Even if the synthesized speech decodes a completely different set of words, it will have high correlation with the target speech as long as the words are synchronized to the target speech. Second, the synthesized speech naturally synchronize with the target speech since the timing of syllables are decoded from the same neural signals for both target and synthesized speech, although in different manners.

We agree with the reviewer that the correlation coefficient between target and synthesized speech is not a sufficient metric to evaluate speech synthesis performance. However, it is useful to track and compare performance across days and tasks (i.e., for coarse internal comparisons). Although the absolute value of correlation coefficient does not give an objective insight on intelligibility, overall higher value indicates better performance. Hence, it has been widely used in the speech BCI literature, following which, we have also used it in this manuscript for easier benchmarking and comparison with previous studies. Indeed, there is a need for a better automated metric to assess synthesis performance, which is still an important open question in the field.

Following reviewer's suggestion to make the correlation coefficient more robust, we have now modified the method by removing the long pauses between the words (i.e., conjoining all words in the synthesized and target speech) before computing the Pearson correlation coefficient. This method is described in Methods section **line 753.** We have also replaced all the reported values of correlation coefficient throughout the manuscript with this new more stringent method, including the results in **Fig 2.b, d, f, i.** Consequently, we have also replaced the results for Mel-cepstral distortion (MCD) in **Extended Data Fig. 3a** by recomputing MCD after removing silences. New analyses done as part of the response to reviewers also use this new scheme for computing correlation coefficient.

Fig. 2 line 197:

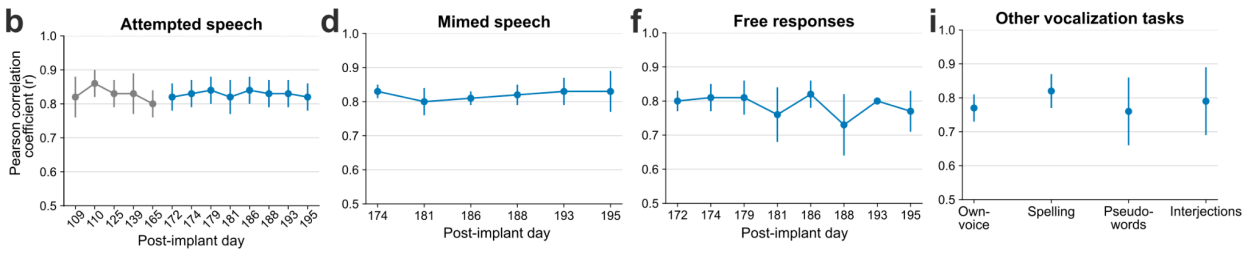
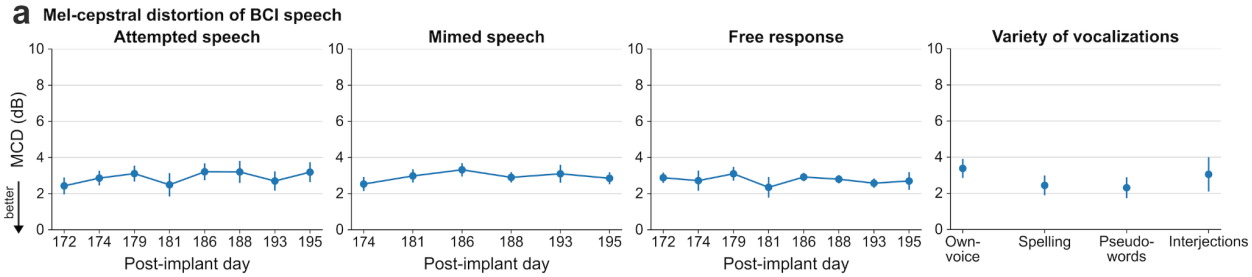


Fig. 2. Voice neuroprosthesis allows a wide range of vocalizations…
b. Pearson correlation coefficients (mean ± s.d) for attempted speech of cued sentences across research sessions excluding silences between words.

Extended Data Fig. 3a line 1030:



Extended Data Fig. 3. Additional BCI speech synthesis performance metrics. a. Mel-cepstral distortion (MCD) is computed across 25 Mel-frequency bands between the closed-loop synthesized speech and the target speech after removing silences between words. The four subpanels show MCDs (mean ± s.d) between the synthesized and target speech for different speech tasks in evaluation research sessions.

Methods line 753:

Evaluation of synthesized speech

Automated evaluation metrics of synthesized speech. We evaluated synthesized speech by measuring the Pearson correlation coefficient between the synthesized and TTS-derived target (fully intelligible) speech after excluding silences between words (since T15 took a pause after each word). Removing silences between words before computing correlation made the measure more sensitive to the quality of synthesis because it would not primarily benefit from just predicting speech versus silence. We computed average Pearson correlations across 40 Mel-frequency bands of audio sampled at 16 kHz. First, the Mel-spectrogram with 40 Mel-frequency bands was computed using sliding (Hanning) windows of 50 ms with 10 ms overlap and converted to decibel units. Then, silences between words were removed from Mel-spectrograms of target and predicted speech independently to conjoin all words. After removing silences, Mel-spectrograms of the target speech and synthesized speech were of different length. We therefore time-aligned the two spectrograms using dynamic time warping and then computed Pearson correlation coefficient between different Mel-frequency bands. We also computed Mel-cepstral distortion (MCD), another common metric used to evaluate speech synthesis models, between the synthesized speech and the target speech using the method described in [28], after removing silences between words.

Results line 136:

The synthesized voice was similar to the target speech (**Fig. 2b**), with a Pearson correlation coefficient of 0.83±0.04 across 40 Mel-frequency bands, after removing silences between words (**Extended Data Fig. 3a** reports Mel-cepstral distortion).

3.b Furthermore, the paper suggests that target speech is synthesized based on the text shown to the patient. However, target speech is also present and is used for evaluation for spontaneous speech. I wonder how the target speech is constructed in these scenarios.

Thank you for noting that our methods were not sufficiently clear about how we construct target speech to evaluate spontaneously synthesized speech. During spontaneous speech trials,

along with brain-to-voice synthesis, we also ran a separate brain-to- text decoder (Card 2024, NEJM) which displayed decoded words on the screen as T15 attempted to speak, based on which T15 later confirmed what he said and corrected the text decoding, ensuring that for each spontaneous speech trial we got the correct ground truth text. We used this ground truth text confirmed by T15 to generate target speech for spontaneous free response trials. We have now clarified this as follows.

Results line 186:

> We used a simultaneous text decoder[1] and confirmed all responses with T15 to get the ground truth text of what he said, which was used to generate target speech and evaluate synthesized speech.

3.c If the target speech is synthesized online during speech production (through instantaneous text decoding and instantaneous TTS), it is an alternative method for speech synthesis.

We thank the reviewer for pointing out that we did not sufficiently explain why this method differs from text-decoding followed by TTS (even if that is done with lower latency). The novelty of the approach used in this study lies in our direct "brain-to-voice" approach for low latency speech synthesis with inference latency of 10 ms (and a total latency of ~30 ms with continuous audio playback) as shown in the **Extended Data Fig. 2.** This is at least an order of magnitude faster than the inherently longer text decoding latency, where first the individual phonemes are predicted, then they are strung together by a language model to form words, and sentences. Even if each word is synthesized with a TTS after it is decoded (without waiting for a longer-context language model), it will incur a delay of ~1s (depending on the length of the word).

Furthermore, TTS audio based on individual words will not have information beyond the phonemic content of the text, such as speed and duration of words being spoken, or intonation information. Our instantaneous synthesis approach not only synthesized voice at the pace of attempted speech, but also allows decoding and modulation of paralinguistic features in the synthesized voice, enabling the user to be more expressive through the BCI-voice.

We have added a sentence to the Introduction **line 64** to better explain this:

> While the aforementioned studies were done with able speakers, a study [3] with a participant with anarthria adapted a text decoding approach to decode discrete speech units acausally (i.e., using neural data from both before and after a given utterance) at the end of the sentence to synthesize speech from a 1,024-word vocabulary. However, this is still very different from healthy speech, where people immediately hear what they are saying and can use this to accomplish communication goals such as interjecting in a conversation. Even lower-latency text-to-speech approaches are acausal and would need to wait for the word to be decoded, which would be substantially slower. In this work, we sought to causally synthesize voice continuously and with low latency from neural activity as the user attempted to speak, which we refer to as "instantaneous" voice synthesis to contrast it with earlier work demonstrating acausal delayed synthesis.

4. Compared with the correlation with target speech, open-set speech recognition is a much more reasonable evaluation method. It can be done either with automatic speech recognizers (ASR) or with human evaluators. Even human evaluation is not costly, a few naïve listeners will be enough. Such a test is critical if the purpose of the work is to enable the patient to interact with others. Similarly, the performance of pitch decoding should be judged by humans (e.g., taking out a pair of syllables from the melody condition and asking human participants to judge which one has a higher pitch). Based on the videos, the melody seems to be quite difficult to judge for myself so that I would like to see ratings from a group of naïve listeners.

We agree with the reviewer that open transcription of synthesized voice by naïve human listeners is a better and more sensitive evaluation metric. As per the reviewer's suggestion, we have had naïve listeners perform open transcription of brain-to-voice synthesized speech as detailed in the **common response #1** on **page 1** of this document. We have added these results to the main results (**Fig. 2l**) in the manuscript. Open transcription yielded phoneme error rate of 34.00% and word error rate of 43.75% for the neurally-synthesized voice. To our knowledge, this is the lowest word error rate achieved by a speech synthesis BCI in the literature. For comparison, T15's residual speech had a word error rate of 96.43%. This shows that the BCI voice is substantially more intelligible than T15's dysarthric speech.

As suggested by the reviewer, we also conducted human evaluation of the neurally-synthesized pitch. We gave human listeners pairs of synthesized pitch notes from the closed-loop three-pitch melody task (which also included the trials in video 12) and asked human listeners to select the note with higher pitch. Each of 189 pitch-pairs were assessed by 7 human listeners. Overall, we found that human pitch classification accuracy was 73.02% .

These new human pitch evaluation results are included in the revised main manuscript **Fig. 3i** (bottom) and **line 307**:

Fig. 3i line 245:

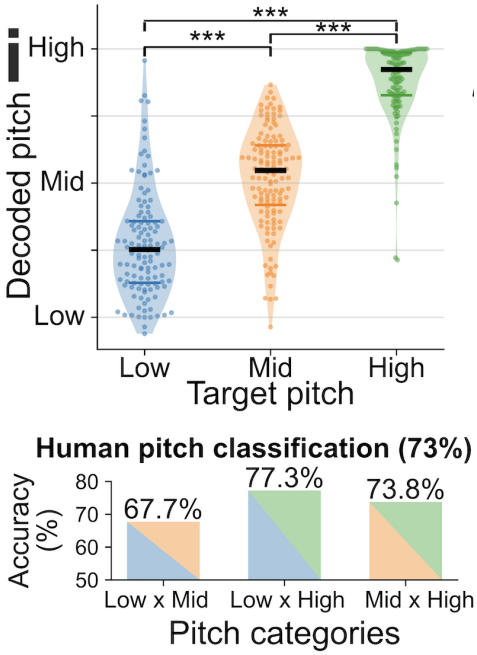


Fig. 3. Modulating paralinguistic features in synthesized voice. …

i. (Top) Violin plots showing significantly different decoded pitch levels for low, medium and high pitch target words (Wilcoxon rank-sum, $p=10^{-14}$ with correction for multiple comparisons, $n_1=122$, $n_2=132$, $n_3=122$). Each point indicates a single trial. (Bottom) Average accuracies of identifying higher pitch from the pairs of synthesized pitch notes (low × mid: $n_1=62$, low × high: $n_2=66$, mid × high: $n_3=61$) by human listeners.

Results line 307:

To assess human perception of synthesized pitch levels, human listeners were asked to select the higher pitched note from 189 pairs (low × mid, low × high, mid × high). Average human classification accuracy of pitch was 73.02% (**Fig. 3i** bottom).

Details of the evaluation method are included in Methods section (**line 806**) as follows:

Human evaluation of synthesized pitch. To assess whether pitch modulation in BCI-synthesized voice was perceptible by human listeners, we assigned human listeners (recruited via Amazon Mechanical Turk) 189 pairs of different pitch notes synthesized by the BCI from randomly selected trials from the closed-loop three-pitch melody singing task. For each pitch pair (low × mid, low × high, and mid × high) drawn from the same evaluation set trial, listeners were asked to select the note with higher pitch. Each pitch pair was evaluated by 7 human listeners and the median choice was selected as the final choice for that pair. Mean accuracy was computed for the three types of pitch pairs.

Minor issues:

5. Terminology: The summary says the study decodes “prosody and intonation”, and the discussion says the study decodes “pitch and intonation”. Notably, pitch, prosody, and intonation are not different things, and it is confusing to list them together. For example, it’s not that the study decodes pitch and something else called intonation that is not based on pitch. Pitch and loudness are critical dimensions for prosody and are decoded in the study. However, the two features are separately decoded in different experiments. As far as I can tell, they are not integrated to form “prosody”. It’s better to say that the study can decode pitch, loudness, and speech rate.

We apologize for the imprecise terminology and appreciate the reviewer pointing this out. We agree with these points and have updated the manuscript text. The Summary (**line 25** onwards) now uses the superset term “prosody” and remove the redundant “emphasize words”:

Brain computer interfaces (BCIs) have the potential to restore communication for people who have lost the ability to speak due to neurological disease or injury. BCIs have been used to translate the neural correlates of attempted speech into text[1–3]. However, text communication fails to capture the nuances of human speech such as prosody~~, intonation~~ and immediately hearing one’s own voice. Here, we demonstrate a “brain-to-voice” neuroprosthesis that instantaneously synthesizes voice with closed-loop audio feedback by decoding neural activity from 256 microelectrodes implanted into the ventral precentral gyrus of a man with amyotrophic lateral sclerosis and severe dysarthria. We overcame the challenge of lacking ground-truth speech for training the neural decoder and were able to accurately synthesize his voice. Along with phonemic content, we were also able to

decode paralinguistic features from intracortical activity, enabling the participant to modulate his BCI-synthesized voice in real-time to change intonation~~, emphasize words~~ and sing short melodies. These results demonstrate the feasibility of enabling people with paralysis to speak intelligibly and expressively through a BCI.

The Results in **line 232** are updated here:

Paralinguistic features such as changes in pitch and cadence play an important role in human speech, allowing us to be more expressive.

The Discussion has been updated using the reviewer's suggested phrasing (**line 834** onwards):

Furthermore, this study demonstrates that a brain-to-voice neuroprosthesis can restore additional communication capabilities over existing brain-to-text BCIs[1–3,28]. Neuronal activity in precentral gyrus encoded both phonemic and paralinguistic features simultaneously. Beyond providing a more immediate way to say words, this system could decode the neural correlates of ~~loudness,~~ pitch and speech rate ~~intonation~~. In online demonstrations, the neuroprosthesis enabled the participant to control a variety of aspects of his instantaneous digital vocalization, including the duration of words, emphasizing specific words in a sentence, ending a sentence as either a statement or a question, and singing three-pitch melodies. This represents a step towards restoring the ability of people living with speech paralysis to regain the full range of expression provided by the human voice.

6. Loudness decoding: The study decoded the loudness of an utterance, and I think a few more analyses are necessary to establish what is decoded is loudness. First, I wonder if the same decoder could determine that the mimed speech is softer or quiet or whether the emphasized words are louder. Second, please show the instantaneous output of the decoder when applied to the imagined melodies, questions, and statements like in Fig. 3. I wonder if aligns with the pitch decoder. If loudness decoding is a part of the conclusion, it should be strengthened. However, I don't think removing this part undermines the study.

We appreciate the reviewer's feedback that the offline loudness decoding results were too brief as presented here and that it might be better to remove this part. Given the previous points that pitch and loudness are often correlated as part of prosody, and due to the large amount of content already in the manuscript (and now the addition of new analyses, supplemental figures and text in response to reviewers' questions), we have taken the reviewer's gentle suggestion and opted to remove these results. Thus, the former offline "*Extended Data Fig. 5*: *Loudness decoding from neural activity*" has been removed in the revised manuscript. We believe this will help better keep focus on the closed-loop results and their accompanying supporting analyses. We hope to follow up in the future with a more narrow-and-deep study specifically exploring the neural correlates of attempted loudness.

7. What is MCD? It's not explained or spelled out.

Mel-cepstral distortion (MCD) is another common method used to assess similarity between two audio waveforms by comparing its Mel-cepstral components. It is used to "objectively" evaluate speech synthesis models (in the sense that computing MCD is automated and deterministic).

We have now spelled out the acronym for MCD when we first use it in the manuscript (**line 138:** ("**Extended Data Fig. 3a** reports Mel-cepstral distortion**"**) and provide a citation to the method of computation **line 764**:

> We also computed Mel-cepstral distortion (MCD), another common metric used to evaluate speech synthesis models, between the synthesized speech and the target speech using the method described in [28], after removing silences between words.

Response to Reviewers on the First Revision

Title: An instantaneous voice synthesis neuroprosthesis
(Manuscript ID: 2024-09-19838A)

Dear Editor and Reviewers,

We are grateful to the reviewers and the editor for their enthusiastic and affirmative response to this study. We are pleased to hear that the reviewers are satisfied with our revised manuscript. We have addressed their remaining concerns in this document and have made the required changes in the manuscript as described below.

Throughout this document, our response to the reviewers is in **blue** text like this**.**

> Relevant copied passages from the manuscript are indented in **black** text, like this. The new text added to the manuscript is highlighted in **red**.

Referee #1 (Remarks to the Author):

I am very impressed by the amount of work and rigor the authors invested into addressing all reviewers' comments. I am a lot more convinced now that the impressive results are indeed based on neural activity. Furthermore, the new intelligibility analysis very realistically reflects the acoustic output. Congratulations on this impressive work.

Despite the higher WER than in the "to-text" approach, I believe this project to represent an important step forward for speech neuroprostheses. By providing immediate feedback and allowing the patient to use intonation and accentuation, the manuscript brings the full expressive power closer to patients.

We are delighted to hear this feedback. We thank the referee for reviewing our manuscript again and are glad that we were able to answer their questions.

Methods are well described and statistics used appropriately. The relevant work is credited and the new findings placed in the existing literature.

Thank you!

The only thing I think is necessary before this work can be published would be to make the new, parallel recording of stethoscope, IMU and Neural data with the appropriate prompts available to the general public so that the fusion and each modality individually can be investigated by the community.

We thank the reviewer for this suggestion. We have made the stethoscope, IMU and neural data with prompts from this control experiment available to the general public in Dryad along with the other study data that we plan to share.

Referee #2 (Remarks to the Author):

I thank the authors for their careful and thoughtful address of my concerns. I am satisfied with the revisions.

We thank the reviewer for their feedback, we are pleased to hear this.

one small confusion remains: regarding the new intelligibility test: you mention 23 female plus 5 male naive listeners but metrics are reported for 7. Perhaps not all 28 did their evaluations in person? or more likely 23 is a typo

We apologize for not making this clear in the Methods. A total of 28 naïve human listeners were involved in the in person open transcription and were divided into four groups of 7 listeners. Each group transcribed 30-38 different BCI-synthesized sentences and 5 sentences for T15's dysarthric speech (since each participant had a limited time to complete the task), totaling 90 random eval sentences, 38 video example sentences and 20 dysarthric speech sentences. Thus, each of the sentence trials reported in Fig. 2l (right) were transcribed by 7 evaluators and different sets of trials were transcribed by different groups of evaluators. The groups of evaluators were formed on the first come, first served basis and the trials were chosen at random for each group.

We have now clarified this in Methods section *Evaluation of synthesized speech*, paragraph 2:

> We recruited 28 naïve listeners from the UC Davis community (native English speakers, 23 female, 5 male, aged 18-42 years) to transcribe each sentence from a representative subset of 90 randomly selected voice synthesis trials from the evaluation set.
> …
> All 28 evaluators were divided into four groups of 7 where each group transcribed 30-38 different BCI-synthesized trials and 5 dysarthric speech trials. Thus, each of the 90 evaluation set trials was transcribed by 7 human evaluators. Additionally, open transcription was also conducted on 38 trials from main **Videos 2**, **3** and **4** to help readers quantitatively assess these examples and contextualize what they hear with the overall metrics. To compare the intelligibility of the BCI synthesized speech with T15's residual dysarthric speech, we also asked the human evaluators to transcribe 20 trials of T15's attempted speech recorded via a microphone during the research blocks.

Referee #3 (Remarks to the Author):

The authors have done a good job revising the manuscript, and have incorporated additional data that can help us to interpret what signals are read out by the speech BCI. I'm largely satisfied with the revision, but two points warrant further discussion in the paper.

We thank the reviewer for this assessment and are glad to hear that our revision satisfactorily answered the comments. We have addressed the remaining comments below.

1. Regarding the speech rate, I fully appreciate that the patient experiences significant difficulty in moving his articulatory muscles. However, in principle, the performance of a BCI should not be constrained by muscular control capabilities, as the signal is directly decoded from brain activity. Therefore, the authors should still discuss why the patient speaks slowly even when it is not necessary to control articulatory muscles: For example, in the future, is it possible to design a speech BCI that does not depend on the patient's ability to overtly articulate? Or, does the patient have difficulty 'imagining' fluent speaking?

The reviewer has raised a good point. We would like to clarify that for this study we targeted the placement of the Utah arrays specifically in the precentral gyrus – the area putatively known to

control speech articulator motor commands – with the goal of decoding these motor commands at the source to drive the speech BCI. Hence, we instructed the participant to try to speak as clearly as he could by *attempting* to make articulatory movements (which was slow for him) to obtain a good signal-to-noise ratio for neural correlates of speech-motor commands for high accuracy decoding. This follows our team's long observation (Vargas-Irwin et al. 2018 [1*]) that, for people with tetraplegia, attempting a hand movement is cognitively and, from a neural recording standpoint, neurophysiologically different activity than imagining the same movement.

In theory, with continued use of the closed-loop BCI, the participant may be able to learn to attempt to speak faster with the BCI resulting in faster speech rates, which is as of yet unknown. This may require a strategy that differs from our standard approach of requesting that the participant attempt the desired action (here, to speak). Ongoing work by our team (Kunz et al. 2025 [2*]) is also investigating whether alternative strategies such as "inner speech" or "imagined speech" (without overtly attempting to speak) can be used for decoding speech. In Kunz et al. 2025 we found that imagined speech behavior results in much less accurate decoding from precentral gyrus, but this is still early days for this strategy. It remains to be seen whether these strategies can achieve equivalent high decoding accuracies required for the BCI to be useful using brain signals from the premotor cortex or whether there are different brain areas more suitable to obtain sufficient signal-to-noise ratio to utilize these covert speech strategies in a BCI.

Finally, we'd note that the true "reason" for the slower speech in ALS is complex, as it can result, in any individual, from a combination of the degeneration of upper motor neurons and lower motor neurons. How this translates to the potentially slower production (or not) of neuronal ensemble activities at the layer III/IV border (or more superficial layers, depending on array placement) in areas 6, 4, and 55b is of great interest, but beyond the scope of our study (or current understanding).

[1] Vargas-Irwin CE, Feldman JM, King B, Simeral JD, Sorice BL, Oakley EM, Cash SS, Eskandar EN, Friehs GM, Hochberg LR, Donoghue JP. Watch, Imagine, Attempt: Motor Cortex Single-Unit Activity Reveals Context-Dependent Movement Encoding in Humans With Tetraplegia. Front Hum Neurosci. 2018; 12:450. DOI: 10.3389/fnhum.2018.00450.*

[2] Kunz E.M., Meschede-Krasa B., Kamdar F., Avansino D., Nason-Tomaszewski S.R., Card N.S., Jacques B., Bechefsky P., Hahn N., Iacobacci C., Hochberg L.R., Brandman D.M., Stavisky S.D., AuYong N., Pandarinath C., Druckmann S., Henderson J.M., Willett F.R., Representation of verbal thought in motor cortex and implications for speech neuroprostheses, Cell, 2025 (in press). bioRxiv: https://doi.org/10.1101/2024.10.04.616375*

We have now emphasized in the Methods that the participant used "attempted speech" strategy and included the possibility of BCIs using these alternative speech strategies in the Discussion as follows.

Discussion paragraph 4:

> Due to T15's condition, his attempted speech was slow. Ongoing work is investigating if alternate strategies such as imagined speech can facilitate speech decoding[29].

Methods section *Experimental paradigm*, paragraph 2:

> The participant performed the following speech tasks using the above trial structure: 1) attempting to speak cued sentences; 2) miming (without vocalizing) cued sentences; …

and 10) singing melodies with different pitch level targets (this task had a prompted reference audio cue for the melody which was played during the delay period). For all the above tasks, the participant was instructed to attempt to speak as clearly as he could i.e. attempt to make articulatory movements (without vocalization for the miming condition). His attempted speech was slow.

2. Output-potent/null space: The output-potent/null space is defined as neural dimensions that can/cannot linearly decode speech features. It is then demonstrated that both spaces can support nonlinear speech decoding. Thus, the distinction between output-potent/null spaces appears to reduce to whether a neural dimension is linearly or nonlinearly related to speech features, which may not be fundamental difference between output-potent and null spaces.

Thank you for pointing out that we should have provided a bit more explanation in the results and methods section to highlight the important temporal distinctions between the output-potent and output-null spaces, both in terms of how they are calculated and the interpretation. We indeed define the output-potent and output-null as orthogonal neural subspaces estimated by linearly decomposing neural activity [1*], with the output-potent dimensions selected as those best capturing the behaviorally correlated neural activity on a moment-by moment basis. The orthogonal complement of this output-potent subspace is then the output-null subspace.

Both of these orthogonal neural subspaces contain speech-related information that can be decoded to a certain extent using *linear* or *non-linear* decoders. For instance, if the neural activity in absence of speech can linearly predict speech features a second later, it will be deemed as output-null space because it doesn't drive the motor output [1*]. Still, the observed behaviour of the output-null activity that increases *preceding* each word and *decays* as the sentence progresses is not given (it could have been similar to the output-potent activity that increases *during* each each word *consistently*, had it been a non-linearly related component of speech producing neural activity). Nor was it by design that the upcoming speech could be decoded quite accurately from output-null activity (especially if using non-linear methods). Hence, the noteworthy distinction between the output-potent and output-null subspaces is not whether they can/cannot linearly or non-linearly decode speech, rather, it is the timing of these neural dynamics.

Regarding linear and non-linear decoding: speech features could not be decoded with high accuracy by a linear decoder from both output-potent and output-null spaces due to its low predictive power for such a complex output. The linear decoder can roughly predict coarser speech features such as the speech envelope from both subspaces, unlike the neural network based non-linear decoders that have more predictive power to synthesize nuanced voice. Hence, while the linear decoder is not useful for synthesizing high-quality voice, it provided us with a transparent and interpretable tool to estimate output-null/output-potent neural subspaces, consistent with prior literature, which would have been murkier with the "black box" non-linear neural network decoders. In the present analyses, we then used a non-linear decoder to synthesize voice from these subspaces to more accurately quantify the amount of speech-related information contained in these neural subspaces.

[1] Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. Nat. Neurosci. 17, 440–448 (2014).*
We have now clarified our wording when describing the definition of output-potent and output-null dimensions in Results, highlighting that we are using an established analysis framework and have added some context in the Methods as follows:

Results section *Output-null neural dynamics of speech*, paragraph 2:

> We estimated the output-null and output-potent neural dimensions by adopting established methods[25,26] for this speech BCI application. These consisted of linearly decomposing the population activity into a subspace with activity that was time-aligned with speech features (output-potent dimensions, which putatively most directly relate to behavioral output) and its orthogonal complement (output-null dimensions, which putatively have less direct effect on the behavioral output). Both subspaces contained substantial speech-related information: the Pearson correlation coefficients of decoding speech using only the output-potent dimensions (which captured 2.5% of total variance) or only the output-null dimensions (97.5% of total variance) were 0.82±0.06 and 0.85±0.07, respectively.

Methods section *Output-null and output-potent activity*, paragraph 1:

> Although the linear decoder did not have sufficient power to decode high-quality speech features, it served as a reliable and interpretable method for decomposing neural activity into orthogonal putatively output-null and output-potent subspaces.