PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES ENGINEERING OPEN ACCESS

# Enhancement of speech-in-noise comprehension through vibrotactile stimulation at the syllabic rate

Pierre Guilleminot[a] and Tobias Reichenbach[a,b,1]

Edited by David Poeppel, New York University, New York, NY; received September 28, 2021; accepted February 14, 2022 by Editorial Board Member David J. Heeger

**Speech unfolds over distinct temporal scales, in particular, those related to the rhythm of phonemes, syllables, and words. When a person listens to continuous speech, the syllabic rhythm is tracked by neural activity in the theta frequency range. The tracking plays a functional role in speech processing: Influencing the theta activity through transcranial current stimulation, for instance, can impact speech perception. The theta-band activity in the auditory cortex can also be modulated through the somatosensory system, but the effect on speech processing has remained unclear. Here, we show that vibrotactile feedback presented at the rate of syllables can modulate and, in fact, enhance the comprehension of a speech signal in background noise. The enhancement occurs when vibrotactile pulses occur at the perceptual center of the syllables, whereas a temporal delay between the vibrotactile signals and the speech stream can lead to a lower level of speech comprehension. We further investigate the neural mechanisms underlying the audiotactile integration through electroencephalographic (EEG) recordings. We find that the audiotactile stimulation modulates the neural response to the speech rhythm, as well as the neural response to the vibrotactile pulses. The modulations of these neural activities reflect the behavioral effects on speech comprehension. Moreover, we demonstrate that speech comprehension can be predicted by particular aspects of the neural responses. Our results evidence a role of vibrotactile information for speech processing and may have applications in future auditory prosthesis.**

audiotactile integration | speech-in-noise comprehension | multisensory processing | EEG

## Significance

Syllables are important building blocks of speech. They occur at a rate between 4 and 8 Hz, corresponding to the theta frequency range of neural activity in the cerebral cortex. When listening to speech, the theta activity becomes aligned to the syllabic rhythm, presumably aiding in parsing a speech signal into distinct syllables. However, this neural activity cannot only be influenced by sound, but also by somatosensory information. Here, we show that the presentation of vibrotactile signals at the syllabic rate can enhance the comprehension of speech in background noise. We further provide evidence that this multisensory enhancement of speech comprehension reflects the multisensory integration of auditory and tactile information in the auditory cortex.

Speech is a highly complex acoustic signal, the comprehension of which necessitates similarly complex hierarchical processes, involving the segmentation and analysis of individual phonemes, syllables, and words to extract semantic content (1–3). Given this complexity, the human ability to understand speech in adverse listening conditions is particularly astounding. We can understand a particular speaker even when background noise is louder than the target speech, when a speech signal is accompanied by multiple reverberations in interior spaces with few acoustic absorbers, and even when perceiving the acoustic signal through a hearing aid or a cochlear implant that provides much less information to the brain than a healthy inner ear (4–7).

Speech comprehension can be supported by other sensory modalities. Lip reading, for instance, refers to the human ability to integrate visual information from the moving lips with the auditory signal to aid speech comprehension (8, 9). Such audiovisual integration can allow humans to understand speech in background noise that is several times louder than they would otherwise be able to cope with. Another example of multisensory speech processing is the addition of tactile signals, delivered through air puffs to the neck of participants, while they hear a single syllable (10, 11). Such aero-tactile stimulation can thereby shape what syllable the participant perceives.

The neural integration of the different sensory streams relies on several processes (12–14). A particularly important temporal mechanism is the neural tracking of rhythms in the multisensory signals (15, 16). The delta frequency band, 1 to 4 Hz, corresponds to the rate of words, while the theta frequency band, 4 to 8 Hz, contains the syllable rate. When a person listens to a speech signal, the neural activity in both frequency bands tracks these speech rhythms (17, 18).

The neural speech tracking reflects acoustic and linguistic processes. Both delta- and theta-band tracking are, for instance, modulated by selective attention to one of several competing talkers (19–21). In addition, the delta-band activity informs on linguistic aspects such as word similarity, surprisal of word sequences, and speech comprehension (22–24). Theta-band activity can allow one to decode acoustic aspects such as the signal-to-noise ratio (SNR) and may allow the brain to parse a speech stream into distinct syllables (24–26). Moreover, altering the theta-band tracking through transcranial alternating-current stimulation can impact speech comprehension, evidencing a functional role in speech processing (27).

Author affiliations: [a]Department of Bioengineering, South Kensington Campus, Imperial College London, London SW7 2BX, United Kingdom; and [b]Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany

Author contributions: P.G. and T.R. designed research; P.G. performed research; P.G. and T.R. analyzed data; and P.G. and T.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.P. is a guest editor invited by the Editorial Board.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: tobias.j.reichenbach@fau.de.

Published March 21, 2022.

Regarding multisensory speech processing, animal studies have demonstrated that visual information can lead to phase resets of auditory cortical activity (28, 29). Such resets of phase relationships related to the cortical tracking of speech rhythms could both enhance and impede speech comprehension, depending on the magnitude and direction of the reset (15, 16).

Recordings of neuronal activity in animals have uncovered that somatosensory stimulation can influence the delta- and theta-band activity in the auditory cortex as well (30–34). In both humans and monkeys, such auditory–somatosensory integration occurs in the primary auditory cortex (34–39). In addition, the latency of the somatosensory response in the primary auditory cortex was found to be only about 12 ms longer than the latency of auditory-evoked activity (32). Somatosensory stimulation can thus provide a mechanism to alter the cortical tracking of speech rhythms and to consequently modulate speech comprehension (34). A recent study paired a speech signal to vibrotactile stimulation with the speech envelope, but did not observe an effect on speech-in-noise comprehension (40). However, this study employed a continuous tactile signal that included both delta- and theta-band portions. Our previous work on transcranial alternating-current stimulation showed that theta-band stimulation, but not delta-band stimulation, led to a modulation of speech-in-noise comprehension (27).

Here, we combined the presentation of speech in noise with sparse vibrotactile pulses that followed the syllabic rhythm in the theta band. We then investigated whether the vibrotactile pulses influenced the comprehension of the speech signal. We also employed electroencephalographic (EEG) recordings to study the effects of the tactile stimulation on the cortical speech processing, as well as to investigate the impact of the acoustic signal on the cortical processing of the tactile pulses. To illuminate the neurobiological mechanisms of the audiotactile speech processing, we finally related the electrophysiological observations to the behavioral findings.

## Results

We presented subjects with speech in background noise paired to vibrotactile stimulation (Fig. 1). The tactile signal was derived from the speech signal: Individual pulses were located at the perceptual centers of the different syllables. To investigate the influence of timing differences, we then considered five different temporal delays between the audio signal and the tactile stimulation. The delays were −180 ms, −120 ms, −60 ms, 0 ms, and 60 ms, in which a positive value meant that the tactile signal preceded the audio one. To control for the stimulation with the tactile pulses per se, and hence for potential placebo effects, we employed two control conditions. First, we used a tactile sham stimulus in which the pulses followed the syllable timings of an unrelated speech stimulus. Second, we presented the volunteers with the audio signal alone. Taken together, we hence considered seven different types of stimulation, in the following referred to as different conditions.

**Behavioral Measures.** We first determined whether the vibrotactile stimulation affected the speech comprehension of the subjects. We therefore presented each volunteer with different, semantically unpredictable sentences in speech-shaped noise. The noise level was chosen such that the subjects understood approximately half of the key words of the sentence. The average rate of syllables in these sentences, and hence the average rate of vibrotactile pulses, was 4.5 Hz (Fig. 2*A*). After each sentence, we asked the subject to repeat what they understood and recorded the percentage of correct key words. In addition, we also asked the volunteers to rate how comfortably they felt they could understand the speech signal.

We found that subjects understood, on average across the seven different conditions, $43 \pm 2\%$ of the key words (mean and SEM). The comprehension was, however, not equal across the different conditions. Instead, we observed that the dependence

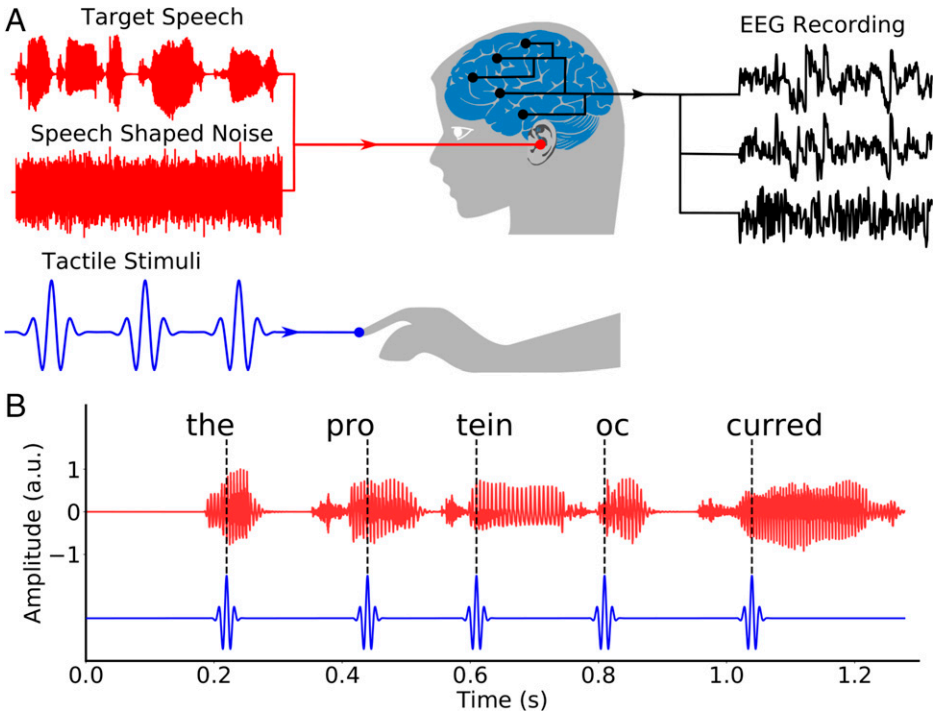


**Fig. 1.** Experimental setup. (*A*) Each subject listened to speech in background noise, while experiencing vibrotactile stimulation at their right index finger. Brain activity was recorded simultaneously through EEG. (*B*) The tactile stimulation consisted of discrete pulses that were located at the perceptual centers of the syllables in the speech stream. A.u., arbitrary units.

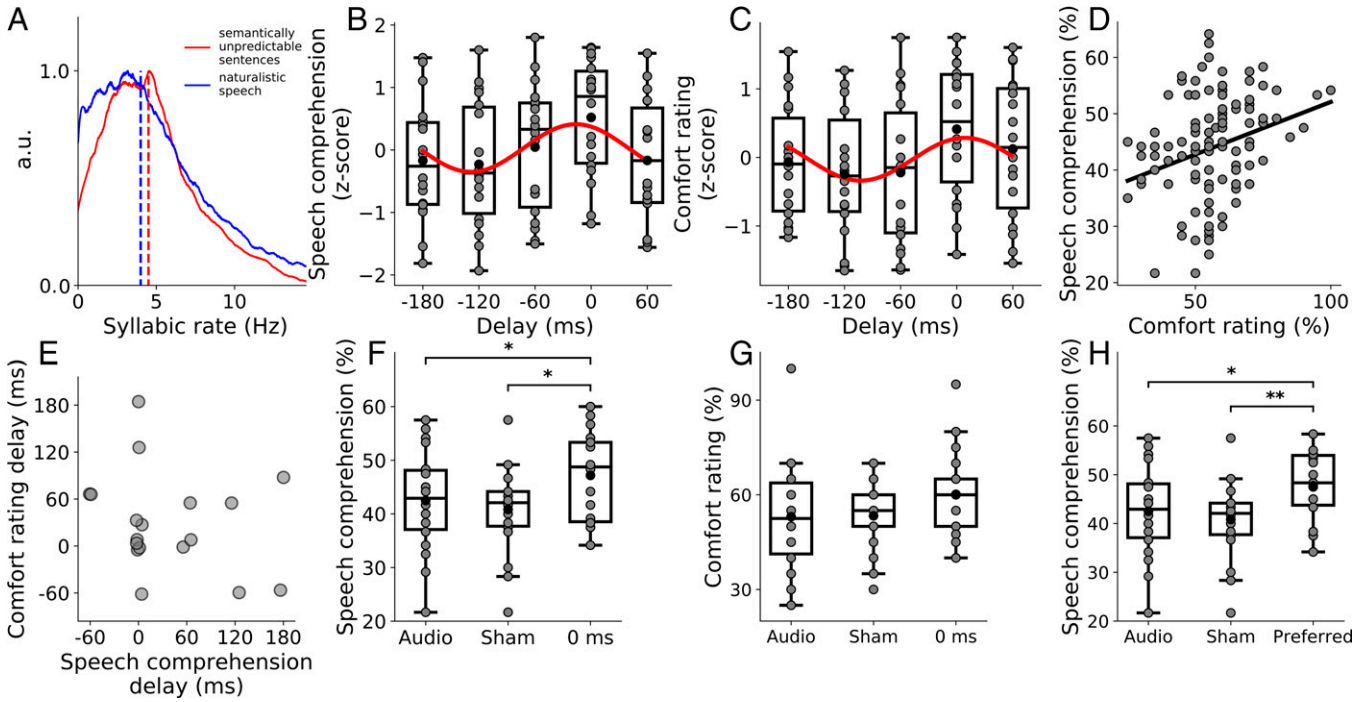


**Fig. 2.** Behavioral results on speech comprehension and syllabic rate. (*A*) Distribution of the syllabic rate of the semantically unpredictable sentences used during the behavioral experiment (red) and of the naturalistic speech used in the electrophysiological recordings (blue). The means are shown as dashed lines. (*B* and *C*) The z-scores of the speech-comprehension scores (*B*) and of the comfort rating (*C*) at the different delays of the tactile stimulation show approximately sinusoidal variation at the rate of the syllables (red lines). The gray disks denote results from individual subjects. (*D*) Comprehension scores are positively correlated (black line) to comfort ratings across the different conditions and subjects (gray disks). (*E*) Comparison between the optimal delays for comfort rating and speech-comprehension subject-wise. The correlation between the two is nonsignificant. (*F* and *G*) Comparisons between the comprehension scores (*F*) and the comfort ratings (*G*) in the auditory-only condition, in the sham condition, and in the condition at which the auditory and tactile streams are synchronized (delay of 0 ms). Data from individual subjects are shown as gray disks. (*H*) Comparison between the comprehension scores for the auditory-only condition, the sham condition, and the condition at which a subject reported the maximal comfort. Gray dots represent data from individual subjects. A.u., arbitrary units.

of the speech-comprehension score on the delay of the tactile stimulation exhibited an approximately sinusoidal variation with a frequency of the syllabic rate of 4.5 Hz ($P = 0.035$, false discovery rate (FDR) correction for multiple comparisons, $R^2 = 0.063$; Fig. 2*B*). The maximum of the sinusoidal fit occurred at a delay of $-17$ ms and hence when the tactile stimulation occurred approximately in synchrony with the speech signal. This was confirmed by an analysis of the distribution of the delays that led to the highest level of speech comprehension for each subject. The distribution was nonuniform ($P = 0.034$), with a peak at 0 ms, and was fitted well by a sinusoidal variation ($P = 0.049$, $R^2 = 0.672$).

To determine whether the tactile stimulation could not only modulate, but also enhance, speech comprehension, we compared the comprehension score at the delay of 0 ms to those obtained from the sham condition, as well as the audio-only stimulation (Fig. 2*F*). A Friedman test revealed significant differences between the three conditions ($P < 0.005$), and subsequent pairwise tests showed that the comprehension score at the delay of 0 ms was significantly higher than in the sham condition ($P = 0.030$, FDR correction for multiple comparisons), as well as in the audio-only condition ($P = 0.027$, FDR correction for multiple comparisons). The audiotactile stimulus without a temporal delay yielded a speech-comprehension score that was $4.7 \pm 2$ % higher than that in the audio-only condition, and $6.3 \pm 2$ % higher than that obtained during the sham stimulation (mean and SEM). No significant difference emerged between the audio-only and the sham condition ($P = 0.32$, FDR correction for multiple comparisons).

The comfort rating exhibited similar behavior. At an average value of $50 \pm 3\%$, it was also uneven across the different delays and was fitted well by a sinusoidal variation ($P = 0.025$, FDR correction for multiple comparisons, $R^2 = 0.045$; Fig. 2*C*). The sinusoidal fit peaked at a delay of 9 ms. The distribution of the best delays across the different volunteers was nonuniform ($P = 0.006$) and had a peak at a delay of 0 ms, but was not fitted well by a sinusoid ($P = 0.09$, $R^2 = 0.57$), as opposed to the speech-comprehension scores. In addition, the comfort rating at the delay of 0 ms was not significantly different from those obtained under the sham stimulation and when presenting the audio signal alone (Fig. 2*G*).

Because the results for the speech-comprehension scores and for the comfort rating appeared so similar, we quantified their dependence further. First, we observed that, across the different conditions and subjects, both measures themselves showed a significant positive correlation ($r = 0.43$, $P = 0.0014$, $R^2 = 0.079$; Fig. 2*D*). We then evaluated if the optimal delays for comfort and speech comprehension were consistent for the individual subjects. Although both measures revealed a cluster of delays around 0 ms, the correlation between the two was not significant ($P = 0.3$, $R^2 = 0.048$; Fig. 2*E*). The best latency for comfort therefore did not vary systematically with the best delay for speech comprehension. To further explore the link between these two measures, we also compared the speech comprehension at the preferred condition—that is, for the condition at which a particular subject reported the highest comfort—to the two control conditions: the sham stimulation and the audio-only stimulation (Fig. 2*H*). We found that these scores differed significantly (Friedman test, $P = 0.0031$). Pairwise tests revealed that the comprehension score in the preferred condition was significantly higher than in the audio-only condition ($P = 0.025$) and in the sham condition ($P = 0.0027$).

**EEG Recordings.** To investigate the neural mechanisms by which the vibrotactile stimulation could modulate and enhance speech comprehension, we measured the subjects' cortical activity while they listened to speech in background noise. For this purpose, we employed long and engaging audiobooks that maintained subjects' attention and that allowed us to obtain EEG data of sufficient length. The audiobooks had an average syllabic rate of 4.0 Hz.

We utilized the same seven conditions that we considered in the behavioral experiment: audiotactile stimulation with five different delays of the tactile signal, one sham stimulation, and an audio-only condition. However, we now added an eighth condition, in which subjects experienced only the tactile stimulation, without hearing a sound.

These different conditions enabled us to compute the multisensory gain associated to the audiotactile conditions. If the brain was not able to integrate the information from the two modalities, then the neural response in the audiotactile conditions would be a simple addition of the individual responses observed in the audio-only and in the tactile-only conditions. By comparing the neural response to an audiotactile stimulation to the summed neural activity elicited by the individual modalities, we could therefore quantify the amount of multisensory integration in the brain.

The subjects' ability to perform the different tasks was monitored in order to exclude subjects that were not paying sufficient attention. Subjects answered $75 \pm 12\%$ of questions related to the content of the stories correctly, and all obtained a score above chance level. When asked to detect occurrences of a specific tactile pattern, $72 \pm 20\%$ of these were found. Only one subject's performance fell below the chance level on this task due to malfunctions of the recording equipment for this specific task.

**Neural Encoding of Multisensory Information.** To determine the neural encoding of the information in the audiotactile signals, we computed two linear regression models, similarly to previous analysis on multisensory integration related to audiovisual speech processing (15, 16, 41). The first model estimated the EEG recordings from the envelope of the speech signal, an important quantity that relates to the speech rhythms, shifted by different temporal latencies. The coefficients of the resulting linear model are referred to as Temporal Response Functions (TRFs) and quantify the contribution of the speech envelope at a particular latency to explain the EEG recording at a particular electrode (Fig. 3*A*). Using these coefficients, we can then estimate the EEG response. By computing the Pearson correlation coefficient between the estimated and the actual EEG response from a subject, we can obtain a reconstruction score.

From the envelope-TRFs associated to the audio-only condition, we identified three latencies at which the speech envelope yielded particularly large contributions—that is, at which the envelope-TRFs peaked. These latencies were 98 ms, 154 ms, and 268 ms. The topographic maps at these latencies showed an approximately symmetric response between the two hemispheres.

We then estimated the multisensory gain. First, we computed unisensory models of the envelope-TRFs obtained in the audio-only and in the tactile-only conditions. These unisensory models were then shifted appropriately and summed to create additive models for each of the multisensory conditions. We then computed the TRF for each audiotactile condition using the corresponding envelopes simultaneously, therefore obtaining a multisensory model for each of the multisensory conditions. We subtracted the additive models from the multisensory models, yielding the multisensory gain related to the speech envelope. We then computed the global field power (GFP) of the multisensory gain across the different electrodes and compared the resulting multisensory gain function for the envelope (envelope-MGF) to that obtained by a null model (Fig. 3*B*).

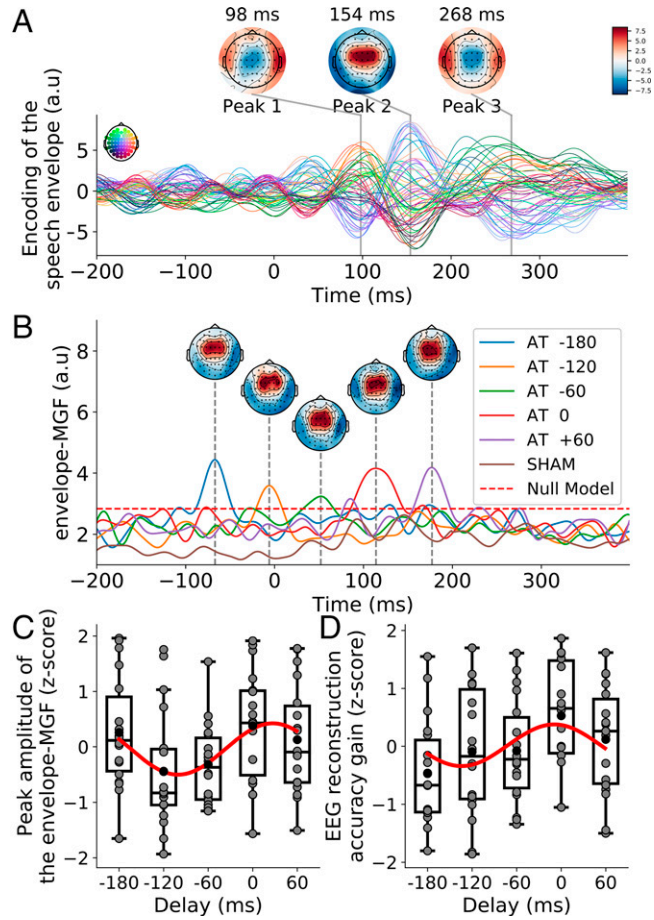


**Fig. 3.** (*A*) The TRF for the unisensory response to the speech envelope shows distinct peaks at positive lags. (*B*) The GFP associated to the multisensory gain of the envelope-TRFs for the audiotactile stimulation, the envelope-MGF, exhibits a single peak at a delay of 114 ms, plus the delay of the tactile signal. The dashed red line indicates the 99% amplitude range of null models. (*C*) After normalization through computing z-scores, the amplitude of the peak of the envelope-MGF depends sinusoidally on the delay of the audiotactile stimulation (red line). (*D*) After normalization through computing z-scores, the multisensory gain in EEG reconstruction depends sinusoidally on the delay of the audiotactile stimulation (red line). A.u., arbitrary units.

The envelope-MGF exhibited one peak for each delay of the tactile stimulation. The peaks occurred at a latency of 114 ms plus the delay of the stimulation. The multisensory integration related to the speech envelope therefore occurred in synchrony with the tactile signal, at a latency of 114 ms. The envelope-MGF associated to the sham stimulation was, in contrast, below the magnitude of the null models and hence statistically insignificant.

The amplitude of the peaks of the envelope-MGF differed between the different delays (Fig. 3*C*). After normalization, we found that the dependence of the peak amplitude on the delay could be described by a sinusoidal variation at the frequency of the syllables ($P = 0.002$, $R^2 = 0.081$; Fig. 3*C*). The sinusoidal fit revealed a peak at the delay of 26 ms. Moreover, the amplitude of the peak at the audiotactile delay of 0 ms was significantly higher than that obtained during sham stimulation ($P < 0.001$).

We then determined the accuracy of the encoding model by computing Pearson's correlation coefficients between the EEG recordings reconstructed from the envelope and the actual EEG signal for EEG channels located in the auditory region of interest (ROI). Across the eight conditions, the correlation coefficients

were $0.012 \pm 0.004$. The distribution of the correlation scores was significantly above chance level ($P = 1e - 13$).

The multisensory gain was evaluated by first computing an additive A+T model that described the sum of the unisensory response to the auditory signal only (A) and that to the tactile signal only (T), as well as by computing a model of the response to the audiotactile stimulation (AT model). The accuracy of the A+T model was then subtracted from that of the AT model, yielding the multisensory gain. The multisensory gains were significantly larger than zero for the delays of –120 ms, –60 ms, 0 ms, and 60 ms ($P = 0.016$, $P = 0.015$, $P = 0.0032$, $P = 0.0096$, respectively; FDR correction for multiple comparisons). The multisensory gains for the sham condition and for the delay of –180 ms were not significantly different from zero ($P = 0.17$, $P = 0.086$, respectively; FDR correction for multiple comparisons).

After normalization, we found that the dependence of the multisensory gain related to the reconstruction accuracy on the delay could be described by a sinusoidal variation at the frequency of the syllables ($P = 0.027$, $R^2 = 0.054$; Fig. 3*D*). The sinusoidal fit revealed a peak at the delay of $-9$ ms. Moreover, the multisensory gain at the audiotactile delay of 0 ms was significantly higher than that obtained during sham stimulation ($P = 0.020$).

As a second model, we similarly computed a linear regression model that estimated the EEG recordings, but this time from spikes located at the centers of the tactile pulses. The resulting pulse-TRFs showed, for the case of the tactile-only condition, three distinct peaks (Fig. 4*A*). These peaks occurred at the latencies of 45 ms, 172 ms, and 241 ms. They were left-lateralized, reflecting the tactile stimulation of the subjects' right hand.

The multisensory gain associated to the pulse-TRFs was computed analogously to that of the envelope-TRFs. We first computed an A+T TRF by shifting and summing the unisensory pulse-TRFs appropriately. We then subtracted the additive pulse-TRFs from the pulse-TRFs that resulted from the audiotactile conditions, yielding the multisensory gain related to the tactile pulses. We then computed the GFP of the multisensory gain across the different electrodes and compared the resulting multisensory gain function for the pulses (pulse-MGF) to that obtained by a null model (Fig. 4*B*). It revealed two peaks in the corresponding pulse-MGF: an early peak at 140 ms and a later peak at 277 ms. The topographies of the two peaks were comparable, symmetric between the two hemispheres, and similar to those of the envelope-TRFs.

The sham stimulation exhibited both significant peaks in the pulse-MGF. To ascertain that the sham stimulation indeed led to a multisensory gain, we computed the difference between the pulse-TRFs of the sham condition and the pulse-TRFs of the tactile-only condition at every latency and channel (Fig. 4*C*). The resulting differences displayed a pronounced peak around a latency of 136 ms, as well as a broad peak around 300 ms—that is, at the delays at which multisensory effects were found in the sham stimulation.

After normalization through computing z-scores, the amplitude of the first peak of the pulse-MGF did not change significantly between the different audiotactile delays ($P = 0.22$, $R^2 = 0.019$; Fig. 4*D*). The amplitude of the smaller second peak, however, exhibited an approximately sinusoidal variation with the delay ($P = 0.022$, $R^2 = 0.056$; Fig. 4*E*). The maximum of the sinusoidal fit occurred at almost the same delay as for the peak of the envelope-MGF, namely, at 25 ms. The second peak for the audiotactile delay of 0 ms was significantly above that obtained for the sham condition ($P = 0.025$).

Analogous to our analysis of the encoding of the speech envelope in the EEG signal, we quantified the accuracy of the encoding model by computing the Pearson's correlation coefficient between the EEG signal reconstructed from the tactile pulses and the actual EEG for the electrodes located in the somatosensory ROI. Across the eight conditions, the correlation score was

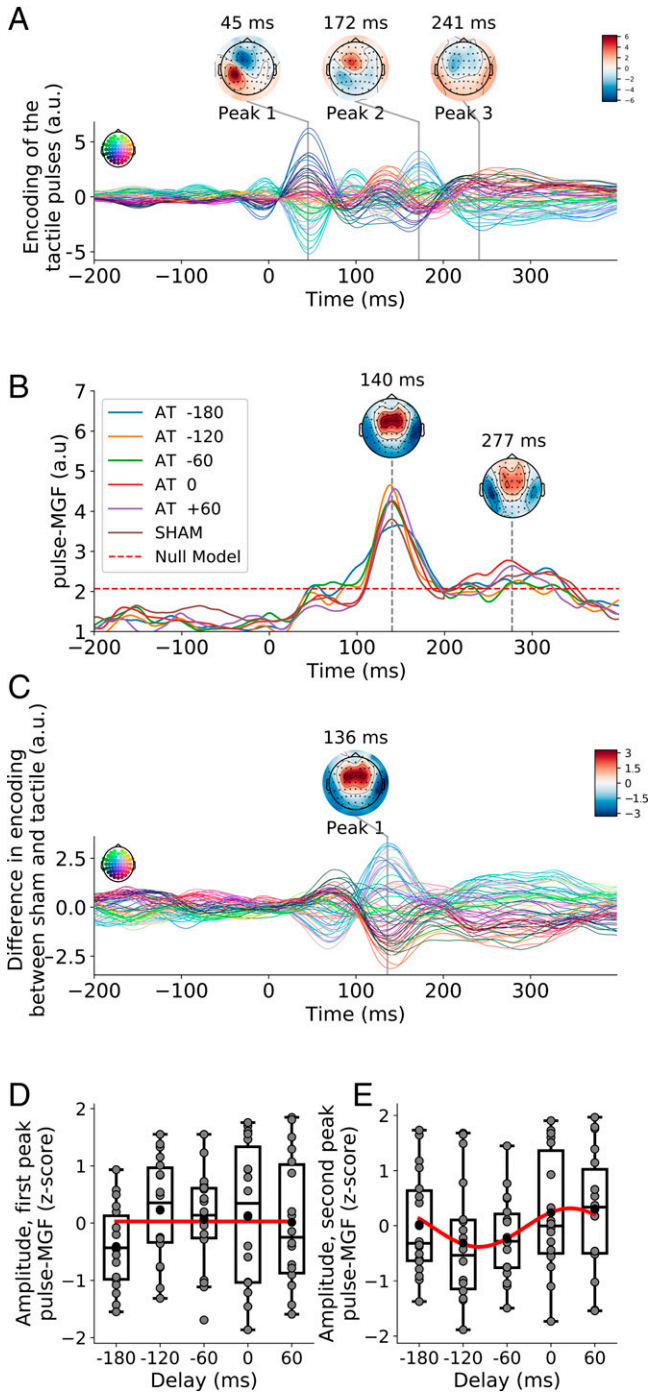


**Fig. 4.** (*A*) The TRF for the unisensory response to the tactile pulses shows distinct peaks at positive lags. (*B*) The GFP of the multisensory gain for the pulse-TRFs, the pulse-MGF, exhibits two peaks, an earlier one at a latency of 140 ms and a later peak at a latency of 277 ms. The topographies are obtained from the audiotactile stimulation without temporal delay. The dashed red line indicates the 99% amplitude range of null models. (*C*) The difference between the TRF to the tactile pulses in the sham condition and the TRF for the unisensory response to the tactile pulses shows a pronounced peak at a delay of 136 ms. (*D*) After normalization through computing z-scores, the amplitude of the earlier peak of the pulse-MGF shows no significant modulation by the tactile delay (red line). (*E*) After normalization through computing z-scores, the later peak of the pulse-MGF exhibits a sinusoidal dependence on the delay of the tactile stimulation (red line). Gray disks denote data from individual subjects. A.u., arbitrary units.

$0.009 \pm 0.003$. Their distribution was significantly above chance level ($p = 1e - 13$).

The multisensory gain was evaluated by subtracting the accuracy of the A+T model from that of the AT model. The multisensory gain, however, was not significantly different from zero in any of the conditions ($p > 0.08$, FDR correction for multiple comparisons). Because of the absence of a supra-additive effect, we did not carry out further analysis.

**Relation between Behavior and the Neural Responses.** As described above, the peak amplitude of the envelope-MGF, the amplitude of the second peak of the pulse-MGF, and the gain in the accuracy of the EEG reconstruction from the speech envelope were modulated by the audiotactile delay. Because these three aspects of the neural encoding displayed a similar sinusoidal modulation with the audiotactile delay as the two behavioral measures, the speech-comprehension score and the comfort rating, we wondered if the behavioral measures could be predicted by the electrophysiological markers. To quantify if such predictions worked, we computed a multiple linear regression model.

We found that the z-scored speech comprehension could not be predicted by that multilinear model ($F = 0.4095$, $P = 0.747$, $R^2 = 0.014$; Fig. 5 *A–C*). The comfort rating, however, was positively correlated with the amplitude of the peak of the envelope-MGF ($F = 3.628$, $P = 0.033$, FDR-corrected, $R^2 = 0.111$; Fig. 5*D*), but not with the two other features (Fig. 5 *E* and *F*).

## Discussion

In this study, we showed that tactile pulses that followed the rhythm of syllables in a speech signal can influence and, in fact, enhance the comprehension of a speech signal in background noise. We further demonstrated that neural activity linked to the speech envelope as well as to the tactile pulses exhibited significant multisensory gains. Moreover, these multisensory gains were modulated by the delay between the tactile pulses and the speech signal in a manner that was very similar to the behavioral data, suggesting that the behavioral effects resulted from multisensory integration of information linked to the rhythms of the speech signal and the tactile stimuli.

To investigate the dependency of speech comprehension as well as of neural activity on the characteristics of the audio-tactile stimuli, we employed different time shifts between the speech signal and the tactile pulses. We found that the dependency of the two behavioral scores that we measured—the speech-comprehension score and the comfort rating—on the tactile delay were both approximately sinusoidal and very similar to each other. The frequency of the sinusoidal variation was that of the syllables in the speech signal. Because the vibrotactile pulses were located at the perceptual centers of the syllables, the rate of the tactile pulses coincided with the rate of the syllables as well.

The highest values of the speech-comprehension measures occurred when the tactile pulses were aligned with the syllables—that is, in the absence of a temporal delay. Speech-in-noise comprehension then improved by 6.3% compared to a sham stimulus. This substantial enhancement suggests that tactile stimuli paired to speech can provide an effective speech-in-noise benefit. Our work hence may open a path to aiding people with hearing impairment to better understand speech in noisy environments, an issue that has remained surprisingly difficult through noise reduction in the audio signal alone (42–46). Such an application would require the extraction of the syllable rhythms of a noisy speech signal, but this problem appears more tractable than extracting the clean speech signal itself and is presumably achievable due to recent progress in deep neural networks for speaker segregation (47).

Other studies have already reported improvement in speech comprehension through tactile stimulation. However, contrary to our work, they either required training, focused on artificially vocoded speech instead of the more natural speech in background noise that we considered, or investigated the recognition of isolated syllables or words (48–50). Moreover, some previous investigations into audio-tactile speech comprehension used more

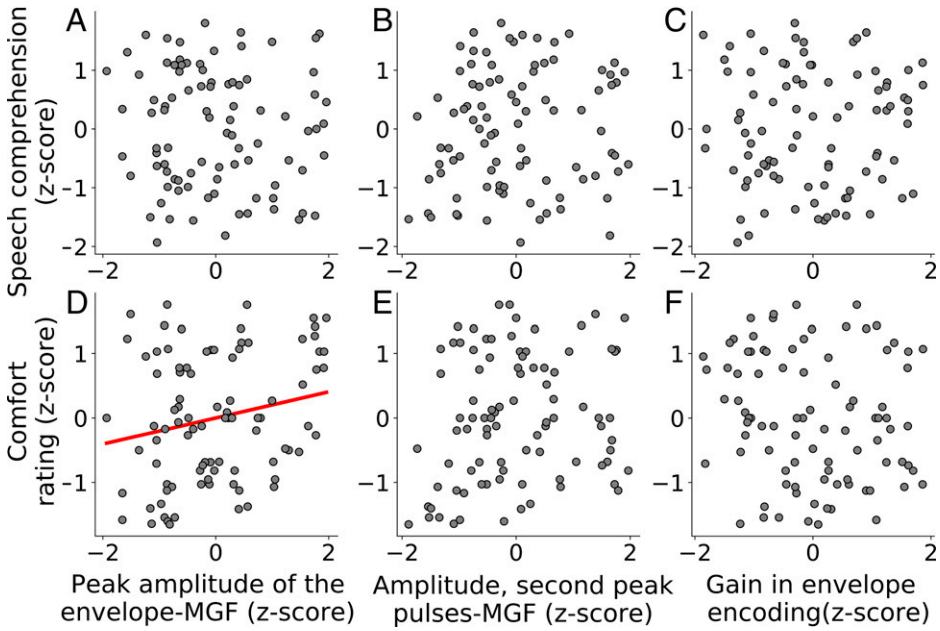


**Fig. 5.** Multilinear regression between behavioral (*y* axis) and electrophysiological (*x* axis) measures. Estimation of the z-scores of speech comprehension, respectively, of the z-scores of the comfort rating from the z-scores of the peak amplitude of the envelope-MGF (*A* and *D*), the amplitude of the second peak of the pulse-MGF (*B* and *E*), and the gain in the accuracy of the EEG reconstruction from the speech envelope (*C* and *F*). The only significant neural predictor of behavior is the peak amplitude of the envelope-MGF: Higher amplitudes signify a higher comfort rating (*D*).

continuous, low-frequency tactile stimuli that replicated either the envelope or the fundamental frequency of speech (40, 48). In contrast, the vibrotactile pulses that we used were sparse and focused on syllable timing rather than on other acoustic properties of speech. The neural integration of the audiotactile stimuli that we considered here appeared to employ existing neural pathways that led to multisensory effects without the use of training or feedback. Further studies, however, are required to further delineate the behavioral and neural correlates of vibrotactile signals that relate to different aspects of a speech signal.

Our working hypothesis was indeed that speech comprehension relies partly on neural tracking of the syllable rhythm in the theta range and that this activity can be modulated through somatosensory stimulation (17, 30, 51). To investigate whether our observed multisensory effects on speech comprehension likely originated from such a neural mechanism, we quantified the neural tracking both of the speech envelope, which contains information on the syllable rhythm, and of the tactile pulses, which followed the perceptual centers of the syllables.

The neural encoding of the speech envelope suggested the presence of sources in the left and right auditory cortex, in accordance with previous studies (41). The neural response to our particular tactile stimuli displayed relatively fast dynamics with a strong early response at 45 ms and a later contribution at 172 ms, both occurring on the contralateral side of the stimuli presentation, to the right hand, and in agreement with earlier work on tactile event-related potentials (52). We note that, due to the lateralization of the neural response to the vibrotactile stimulation, our findings do not necessarily generalize directly to stimulus presentation to the left hand or to left-handed subjects (53).

For both the neural response to the speech envelope and the neural response to the tactile pulses, we found significant multisensory gains (54). The gain associated to the speech envelope showed significant contributions at a delay of 114 ms, plus the delay of the tactile pulses. The corresponding scalp topography showed a bilateral response. This bilateral spatiotemporal pattern of the neural response suggests an origin in the auditory cortex, rather than in a somatosensory area, for which activity is localized on the contralateral side of the stimulation (right hand) (30, 55).

The multisensory gain associated to the neural response to the tactile pulses revealed two main contributions, both of which were different in latency and topography from the neural responses to the tactile pulses alone. The early contribution occurred at a delay of 140 ms and the later at a latency of 277 ms. The scalp topographies exhibited similar activity in both hemispheres and suggested an origin in the auditory cortex.

Moreover, the sham condition elicited a multisensory response at both peak latencies. This finding suggests that tactile pulses are integrated differently when subjects are attending speech, as compared to when they are attending the vibrotactile pulses, even when the pulses are not coherent with the auditory stream. Indeed, in the tactile-only condition, subjects focus their attention on the pulses, while their attentional focus is on the audio signals in all other conditions, including the sham stimulation. As a further corroboration of this observation, we found a significant difference between the TRFs obtained in the sham condition and those of the tactile-only condition, with a pronounced peak at a delay of 136 ms and a broad peak around 300 ms.

However, while supra-additivity could be observed in the coefficients of the TRFs, this did not translate into a significant multisensory gain in the correlation score between the EEG signal reconstructed from the tactile pulses and the actual EEG recording. This might be explained by noise and the necessity to correct for multiple tests.

The multisensory gain to the speech envelope in both the parameters and reconstruction score as well as the later contribution to the multisensory gain to the tactile pulses were modulated by the delay of the tactile stimuli. The resulting dependencies were very similar to those of the behavioral measures: These multisensory gains displayed a sinusoidal dependency on the delay of the tactile pulses. The maximal gains occurred at a delay of ∼25 ms, which was comparable to, although slightly later than, the delays of −17 ms and 9 ms at which the behavioral scores peaked, respectively. This similarity between the dependencies of the neural multisensory measures and the behavioral ones suggested that the effect on speech comprehension did indeed result from the multisensory integration of the rhythmic information in the audiotactile stimuli.

The audiotactile modulations of these behavioral and electrophysiological responses could be attributed to a phase-reset mechanism (34). Somatosensory input can presumably reset the phase of slow neural oscillations in the primary auditory cortex. Because these oscillations appear to play a role in syllable parsing, the vibrotactile signal may modulate syllable parsing, and hence speech comprehension, in a manner that depends on the phase shift between the vibrotactile signal and the auditory one. This hypothesis is supported by our findings, as well as by our observation that the multisensory integration presumably originates in the auditory cortex. Moreover, we found that the optimal lag between auditory and tactile streams was around 0 ms. Such a small delay is coherent with previous research on macaques, where the somatosensory-evoked activity in the early auditory cortex lagged the auditory-evoked activity by about 12 ms only (32).

The similarity of the dependency of the neural multisensory gains and the behavioral scores allowed us to infer the comfort rating from the multisensory gain related to the speech envelope. Both quantities displayed a positive correlation, evidencing that a higher multisensory gain was associated to a higher comfort rating.

However, the speech-comprehension score could not be predicted from the neural measures. This could be due to several reasons. First, the behavioral measure refers to a high linguistic level, whereas the neural measures relate to low-level acoustic aspects, resulting in a mismatch in the levels of speech processing. Second, the behavioral and the electrophysiological data were collected on two different tasks and on different days. The behavioral data assessed the participant's repetition of semantically unpredictable sentences in noise, which involved more short-term memory than the naturalistic speech comprehension during which the neural data were recorded. Third, there is a small difference in the delays at which the sinusoidal measures reached their peaks. Although these differences in the delays represented only a fraction of a cycle of the sinusoidal variation, they could already suffice to impair a direct correlation.

An earlier investigation into behavioral and neural measures of audiotactile speech comprehension employed tactile signals that followed the envelope of the speech stimuli (40). Although that study found, similar to our work, significant multisensory gain in the neural response to the speech envelope, it did not identify a behavioral effect. Because we employed tactile pulses at the perceptual centers of the syllables, related to the theta rhythms, this suggests that it is indeed the theta portion of the speech rhythms, rather than, e.g., the slower word-level delta rhythms, that can aid speech comprehension. Future investigations of different types

and rhythms of tactile stimuli derived from speech rhythms will allow for further clarification of the neural mechanisms of audiotactile speech perception, as well as to further improve the efficacy of such stimuli for enhancing the comprehension of speech in noise.

## Materials and Methods

**Participants.** Nineteen young adults (19 to 24 y old, 10 females) participated in the experiment. All volunteers were native English speakers, identified as right-handed, and had no history of neurological disorders or hearing impairment. They did not receive prior training regarding tactile or audiotactile stimulation. Subjects gave their written informed consent before the experiment. The research was approved by the Imperial College Research Ethics Committee. One male participant did not complete the study due to personal reasons.

**Hardware.** Acoustic and vibrotactile stimuli were generated digitally. Both signals were synchronized and converted into analog signals through the RX8 Multi-I/O Processor device (Tucker-Davis Technologies) at the sampling rate of 39,062.5 Hz. The acoustic stimuli were presented to the subjects through insert earphones (ER-2, Etymotic Research) placed in the subject's ear canals. Tactile stimuli were delivered through a vibrating motor (Tactuator MM3C-HF, TactileLabs) that volunteers held between the thumb, index finger, and middle finger of their right hand.

The subject's spoken response was recorded with a microphone (Blue Snowball, BlueDesigns). EEG signals were acquired by using 64 active electrodes (actiCAP, BrainProducts) and a multichannel EEG amplifier (actiCHamp, BrainProducts). The acoustic signals presented to the volunteers were recorded as well in conjunction with the EEG signals (StimTrak, BrainProducts).

**Acoustic Stimuli.** To assess speech comprehension, we used single sentences presented in speech-shaped noise at an SNR of –2 dB and an intensity of 75-dB sound pressure level (SPL), a comfortable intensity for the subjects. The sentences were semantically unpredictable and were generated by using Python's Natural Language Toolkit (56, 57). Each sentence contained four key words. The sentences were converted to audio by using the TextAloud software with a female voice at a sampling rate of 44,100 Hz. Scores were graded by hand twice, at the time of experiment and through the audio recordings.

When recording EEG, we used nonrepeating continuous stories in speech-shaped noise. The SNR and the sound intensity were the same as for the assessment of speech comprehension: We employed an SNR of –2 dB and an intensity of 75 dB SPL. The stories were extracted from "The Children of Odin: The Book of Northern Myths," by Padraic Colum, a publicly available audiobook read by a female speaker, Elizabeth Klett. The narratives were cut into segments of a duration of 2 min and 30 s.

The speech-shaped noise was generated by determining the Fourier transform of the speech material from the sentences or from the audiobooks, so that it presented the same spectral content as the target speech. The syllable rate was evaluated for both types of stimuli by computing the average duration between the vowels of two consecutive syllables, excluding pauses.

**Tactile Stimuli.** We constructed vibrotactile stimuli in which pulses were aligned to the perceptual centers of the syllables in a speech signal. We therefore first extracted the timing of phonemes from the acoustic signals and the accompanying text using the Montreal Forced Aligner (58). Because the continuous stories contained a few unusual words, we used CMU Sphinx-4 to implement a grapheme-to-phoneme conversion and add their pronunciation to the forced aligner dictionary. The onset of the vowels were considered to represent the perceptual center of the corresponding syllables. The obtained timings were checked manually by using Praat (59).

For the pulse at the perceptual center of a syllable, we used a real Morlet wavelet $\psi(t)$:

$$\psi(t) = \psi_0 \sin(2\pi f t) e^{-\frac{t^2}{2\sigma^2}}, \quad \textbf{[1]}$$

in which $t$ denotes time. We considered an amplitude $\psi_0$ of 1.4 V, a carrier frequency $f$ of 80 Hz, and a width $\sigma$ of 7.5 ms. We thus obtained a series of pulses aligned with the syllabic phonetic features for both the individual sentences and the continuous stories.

**Experimental Design.** We considered seven different types of audiotactile stimulation for the behavioral assessments. Five of these types were tactile signals paired to speech at different delays: −180 ms, −120 ms, −60 ms, 0 ms, and 60 ms. A positive delay referred to the tactile signal proceeding the audio stream. As a control condition, we employed a sham stimulation, in which we paired the speech signal to a vibrotactile stimulation that was obtained from an unrelated speech stream. As a second control, we also considered an audio-only stimulation in which no tactile signal was included.

For the EEG recordings, we employed the seven types of audiotactile stimulation that we used for the behavioral investigations. We also added an eighth type, namely, a tactile-only stimulation that did not include an audio signal. The purpose of this eighth type was to allow the quantification of the multisensory gain in the neural response.

For each subject, the assessments were split into three distinct parts that were carried out on three different days. During the experiments, subjects sat in a dimly lit anechoic chamber and were asked to look at a fixation cross.

The behavioral measures of speech comprehension were obtained in the first part. Subjects heard short sentences in noise. After each sentence, they were asked to repeat what they had heard as accurately as possible. They were also asked to rate how comfortably they could understand the target speaker.

Sentences were presented in 16 blocks. Each block contained 15 sentences that all occurred in the same type of audiotactile stimulation. In addition, two blocks were presented at the start to accustom subjects to the task and the stimuli. In those blocks, the type of stimulation was randomized for each sentence. Each type of audiotactile stimulation occurred then twice in the remaining 16 blocks, in a random order.

The speech-comprehension score for each type of audiotactile stimulation was computed as the average percentage of correctly understood keywords and the comfort rating as the average rating for that stimulation type.

The second and third parts of the experiment were dedicated to the EEG recordings. The subjects thereby listened to continuous audiobooks. The speech material was divided into 32 segments, each with a duration of 2 min and 30 s. The electrophysiological recordings were split into two sessions, so that each session would last less than an hour and participants would not be subjected to fatigue. The type of audiotactile stimulation varied randomly from segment to segment, in such a way that each type appeared 4 times in the 32 segments that each participant heard, resulting in 10 min of presentation per type. At the beginning of each segment, a summary of the previous one was displayed on the screen, so that the volunteer did not miss the main information, even if their speech comprehension might have been low.

For the audio-only and the audiotactile stimulation types, after each segment, the volunteers had to answer multiple-choice questions about the speech to which they had just listened. For the tactile-only type, during the stimulus presentation, they were asked to focus on the tactile pulses and detect a specific pattern that was presented to them prior to the experiment.

The orders of the stories, sentences, and conditions were randomized individually for each subject, and all experiments were conducted in a double-blind manner.

**Preprocessing of EEG Recordings.** EEG data were collected at a sampling rate of 1,000 Hz with 64 electrodes referenced to the vertex (Cz). The maximal impedance was kept under 10 $k\Omega$. EEG preprocessing analysis were carried out by using the MNE Python package (60). Preprocessing involved interpolating bad channels, band-pass filtering between 0.1 Hz and 32 Hz (one-pass, zero-phase, noncausal finite impulse response [FIR] bandpass filter of order 33,000), robust detrending (61), rereferencing to the channel average, and artifact removal through independent component analysis (62). A low-pass filter at 20 Hz was applied at the end (one-pass, zero-phase, noncausal FIR bandpass filter of order 660).

**Sensory Features.** We determined the electrophysiological correlates of two distinct sensory features that were designed to capture the rhythmic nature of the stimuli. First, regarding the auditory signal, we considered the envelope of the speech signal. The envelope $e(t_n)$ at time step $t_n$ was obtained through the Hilbert transform of the speech stream and subsequent low-pass filtering at 20 Hz (one-pass, zero-phase, noncausal FIR lowpass filter of order 660, corner frequency of 22.5 Hz).

Second, to characterize the neural processing of the tactile pulses, we considered a feature that measured the center of the pulses. This feature $p(t_n)$ was accordingly one at the center of a pulse and zero at all other times.

For the subsequent analysis, we then combined the two features to the feature vector $\vec{y}(t_n) = (e(t_n), p(t_n))$.

**Forward Model.** To investigate the encoding of the multisensory information in the neural response, we related the two sensory features to the EEG data through a forward model. The latter was a linear spatiotemporal reconstruction of the normalized EEG signal $\{x_i(t_n)\}_{i=1}^{N}$ for $N = 64$ channels at each time step $t_n$ from the two normalized sensory features $\vec{y}(t_n - \tau_k)$ shifted by $T$ different latencies $\{\tau_k\}_{k=1}^{T}$:

$$\hat{x}_i(t_n) = \sum_{k=1}^{T} \alpha_{ij}(\tau_k) y_j(t_n - \tau_k), \quad \textbf{[2]}$$

in which $\hat{x}_i(t_n)$ is the estimate of the EEG signal at channel $i$ and time step $t_n$. The coefficients $\alpha_{ij}(\tau_k)$ of this linear filter describe the neural encoding of the sensory information and constitute the TRF (41).

We considered an equally spaced range of latencies $\{\tau_k\}_{k=1}^{T}$ from $-1,000$ ms to 1,000 ms, resulting in $T = 2,000$ latencies at a sampling frequency of 1,000 Hz. The TRF was estimated for each subject separately by using regularized ridge regression: $\alpha = (Y_j^t Y_j + \gamma I)^{-1} Y_j^t x_i$, with $Y_j$ denoting the design matrix of the $j$th feature of size $n \times T$, with $n$ denoting the number of samples, $I$ the identity matrix, and $\gamma$ the regularization parameter that was fixed at $\gamma = 1$.

To study the multisensory gain of the obtained neural encoding, we considered a supra-additivity criterion (54). We thereby evaluated whether the multisensory response was different from the sum of the two unisensory responses–that is, those obtained in the audio-only and in the tactile-only stimulation. To this end, we summed the TRFs of the unisensory conditions to obtain an additive TRF, $\alpha^{(A+T)}$, that we refer to as the A+T model. The unisensory TRFs were thereby shifted according to the delay between speech and tactile pulses in the corresponding audiotactile stimulus to respect the alignment between the actual stimuli and the neural response. The multisensory gain $\alpha_{ij}^{(gain)}(\tau_k)$ then followed as the difference in the TRF $\alpha^{(AT)}$ for the audiotactile stimulation and the additive TRF of the A+T model:

$$\alpha_{ij}^{(gain)}(\tau_k) = \alpha_{ij}^{(AT)}(\tau_k) - \alpha_{ij}^{(A+T)}(\tau_k). \quad \textbf{[3]}$$

For the sham stimulus, we created an A+T sham model. For each sensory feature and unisensory condition, we first computed the TRFs of the response to a unisensory-unrelated stimulus. These unisensory-unrelated TRFs represented the noise associated to the unisensory TRFs. For each sensory feature, we therefore summed the unisensory-related TRFs related to one sensory modality (auditory or somatosensory) and the unisensory-unrelated TRFs to the other sensory modality. We then assessed the multisensory gain by comparing the TRFs $\alpha^{(AT)}$ for the audiotactile stimulation to those of the additive sham TRFs. In particular, when analyzing the neural response to the speech envelope, we compared the TRFs $\alpha^{(AT)}$ to the sum of the TRFs in the audio-only condition and the TRFs in response to an unrelated stimuli in the tactile-only condition. The multisensory gain in the neural response to the tactile pulses was computed by comparing the TRFs $\alpha^{(AT)}$ to the sum of the TRFs in the tactile-only condition and the TRFs in response to an unrelated stimuli in the auditory-only condition. We chose this approach because the mismatch of the auditory and the tactile signal in the sham stimulus prevented the computation of a meaningful additive model. However, adding these unrelated unisensory TRFs allowed us to maintain the same noise level as with other multisensory gains and enabled a meaningful comparison.

The GFP for the $j$th stimulus feature followed, for each latency $\tau_k$, as the average of the multisensory gain $\alpha_{ij}^{(gain)}(\tau_k)$ over the different EEG channels $i$.

To assess the statistical significance of the GFP, we constructed a null distribution using the latencies before $-200$ ms and after 800 ms. We then set a threshold for statistical significance at the 99th percentile of the null distribution. For each peak of the GFP, and for each of the two sensory features, we then determined if it was above the chance level, and in that case, extracted the peak amplitude for each subject.

We then employed the forward models to assess how well the EEG responses could be reconstructed (16, 41). For the prediction of the EEG recordings from the speech envelope, we used an equally spaced range of latencies $\{\tau_k\}_{k=1}^{T}$ from $-200$ ms to 400 ms, resulting in $T = 600$ latencies at a sampling frequency of 1,000 Hz. For the reconstruction of the EEG signals from the tactile pulses, we used an equally spaced range of latencies $\{\tau_k\}_{k=1}^{T}$ from 0 ms to 400 ms, resulting in $T = 400$ latencies at a sampling frequency of 1,000 Hz. The range of latencies was shorter than when evaluating the TRFs since we did not use a portion of the delays to compute null models. We focused on auditory and somatosensory regions of interest, corresponding to the unisensory responses we observed, and therefore predicted the responses of specific EEG channels only. As the auditory ROI, we considered the channels FC5, FC3, C5, C3, FC6, FC4, C4, and C6, showing a bilateral pattern. As the somatosensory ROI, we considered the channels FC1, FC3, F3, F1, CP5, CP3, P5, and P3 on the contralateral side to the stimulation (right hand). We used the same regularization parameter $\gamma = 1$ as for the computation of the TRFs.

The models were evaluated separately for each subject by using 20-fold cross-validation, resulting in folds of around 30 s. The performance of each model was measured by computing the Pearson's correlation coefficient between the estimated EEG signals and the actual ones. These values were then averaged across the considered ROIs. The empirical chance level for the reconstruction accuracy was calculated by estimating EEG signals from the time-reversed sensory features and computing the Pearson's correlation coefficient between the reconstructed EEG recording and the actual one. We then compared the accuracy of the EEG reconstruction to the chance level using a $t$ test with a significance criterion of $\alpha = 0.05$.

To determine the multisensory gain, as for the evaluation of the TRFs, we considered a supra-additivity criterion (54). For each sensory feature and subject, we computed TRFs for the two unisensory responses and summed the two to create an additive A+T TRF. We thereby shifted the unisensory decoders with respect to each other by the delay of the corresponding audiotactile stimulation. For each sensory feature, the A+T sham model was built as the sum of the unisensory TRF associated to that feature and a TRF between the other sensory feature and an unrelated EEG response.

We then used the additive TRF on the multisensory EEG data to compute Pearson's correlation coefficients $\rho_j^{(A+T)}$ between the estimated and the actual EEG responses. These reconstruction accuracies constituted the baseline of additive audio-tactile integration. The A+T reconstruction accuracies $\rho_j^{(A+T)}$ were then subtracted from the AT reconstruction accuracies $\rho_j^{AT}$ that were obtained from audiotactile stimulation, yielding the multisensory gain at each delay for each subject:

$$\rho_j^{(gain)} = \rho_j^{(AT)} - \rho_j^{(A+T)}. \quad \textbf{[4]}$$

**Statistical Testing.** We first analyzed the dependency of the speech-comprehension score, the comfort rating, the amplitude of the peaks in the neural data, and the EEG reconstruction scores on the different temporal delays. To account for differences between subjects, z-scores of the data were computed. To this end, for each subject, we subtracted the mean and SD of the respective measure across all conditions. Therefore, for each subject, the resulting z-scores were centered around zero and had an SD of one. While the consideration of the z-scores did not impact the nonparametric statistical testing between the different conditions, the z-scores allowed us to reduce intersubject differences, and hence to better assess differences between the various conditions.

The speech stimuli displayed syllabic rates that peaked around the average rates (Fig. 2*A*). Because we hypothesized that the vibrotactile stimuli would influence the behavioral and neural measures through the timing of the pulses in relation to the timing of the syllables in the speech signal, we modeled the temporal dependencies of the z-scores through a sinusoidal variation at the respective syllabic rate. Amplitude and phase of the sinusoidal variation were free parameters and were fitted through least-squares regression. We transformed the exogenous matrix according to a sine wave with this phase and frequency, but an amplitude of one. This allowed us to transform the data so that the relationship between the exogenous and the endogenous matrix was linear and allowed for standard linear regression (63). The statistical significance of the sinusoidal fit could then be assessed though the linear regression with the new modified exogenous matrix. We used a significance criterion of $\alpha = 0.05$.

When the modulation of a behavioral or neural score was statistically significant, we then further assessed how the maximal score across the different delays compared to a baseline score. For speech comprehension and comfort rating, we compared the maximal scores to those obtained in the audio-only and in the sham condition. For the electrophysiological measures, we compared the maximal scores to those obtained from the corresponding additive A+T model, as well as to those obtained from the AT model that described the sham stimulation.

The comparison of the maximal scores to the baseline scores was done through a second-level group analysis using nonparametric statistical tests for repeated measures (Wilcoxon and Friedman). When performing multiple statistical tests, the *P* values were corrected by using FDR correction (64).

As an additional test of the modulation of the speech-comprehension scores and the comfort rating on the delay of the tactile signal, we also computed a scatter plot of the optimal delays for each of these two measures. We also computed the correlation between these optimal delays.

We also assessed whether the speech-comprehension scores in the preferred condition for each subject–that is, in the condition in which they reported the highest comfort rating–compared to the scores in the audio-only stimulation, as well as during sham stimulation. If a subject had identical highest comfort ratings in more than one condition, we averaged the speech-comprehension scores across these conditions.

We employed multiple linear regression to estimate the behavioral measures from the multisensory electrophysiological features. We corrected for multiple comparisons by controlling the FDR via knockoffs at 1,000 iterations. Contrary to other methods of FDR correction for multiple regression, this method accounts for the correlation structure of the exogenous matrix and can achieve an exact FDR correction (65).

**Data Availability.** Data is available in a publicly accessible database (https://zenodo.org/record/5512578) (66).

**ACKNOWLEDGMENTS.** This research was supported by Engineering and Physical Sciences Research Council (EPSRC) Grant EP/R032602/1 as well as by the EPSRC Centre for Doctoral Training in Neurotechnology for Life and Health. We are grateful to the Imperial College High Performance Computing Service (doi: 10.14469/hpc/2232).

1. S. K. Scott, I. S. Johnsrude, The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* **26**, 100–107 (2003).
2. G. Hickok, D. Poeppel, The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
3. I. DeWitt, J. P. Rauschecker, Phoneme and word recognition in the auditory ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E505–E514 (2012).
4. S. D. Soli, L. L. Wong, Assessment of speech intelligibility in noise with the Hearing in Noise Test. *Int. J. Audiol.* **47**, 356–361 (2008).
5. R. W. Hutcherson, D. D. Dirks, D. E. Morgan, Evaluation of the speech perception in noise (SPIN) test. *Otolaryngol. Head Neck Surg.* **87**, 239–245 (1979).
6. K. S. Helfer, L. A. Wilber, Hearing loss, aging, and speech perception in reverberation and noise. *J. Speech Hear. Res.* **33**, 149–155 (1990).
7. M. Armstrong, P. Pegg, C. James, P. Blamey, Speech perception in noise with implant and hearing aid. *Am. J. Otol.* **18**, S140–S141 (1997).
8. W. H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).
9. A. Alsius, R. V. Wayne, M. Paré, K. G. Munhall, High visual resolution matters in audiovisual speech perception, but only for some. *Atten. Percept. Psychophys.* **78**, 1472–1487 (2016).
10. B. Gick, D. Derrick, Aero-tactile integration in speech perception. *Nature* **462**, 502–504 (2009).
11. B. Gick, Y. Ikegami, D. Derrick, The temporal window of audio-tactile integration in speech perception. *J. Acoust. Soc. Am.* **128**, EL342–EL346 (2010).
12. L. Cappelletta, N. Harte, "Phoneme-to-viseme mapping for visual speech recognition" in *ICPRAM (2)* (Citeseer, 2012), pp. 322–329.
13. A. E. O'Sullivan, M. J. Crosse, G. M. D. Liberto, A. de Cheveigné, E. C. Lalor, Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. *J. Neurosci.* **41**, 4991–5003 (2021).
14. J. Tuomainen, T. S. Andersen, K. Tiippana, M. Sams, Audio-visual speech perception is special. *Cognition* **96**, B13–B22 (2005).
15. M. J. Crosse, G. M. Di Liberto, E. C. Lalor, Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *J. Neurosci.* **36**, 9888–9895 (2016).
16. M. J. Crosse, J. S. Butler, E. C. Lalor, Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* **35**, 14195–14204 (2015).
17. A. L. Giraud, D. Poeppel, Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
18. N. Ding, J. Z. Simon, Cortical entrainment to continuous speech: Functional roles and interpretations. *Front. Hum. Neurosci.* **8**, 311 (2014).
19. N. Ding, J. Z. Simon, Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 11854–11859 (2012).
20. C. Horton, M. D'Zmura, R. Srinivasan, Suppression of competing speech through entrainment of cortical oscillations. *J. Neurophysiol.* **109**, 3082–3093 (2013).
21. J. A. O'Sullivan *et al.*, Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* **25**, 1697–1706 (2015).
22. M. P. Broderick, A. J. Anderson, E. C. Lalor, Semantic context enhances the early auditory encoding of natural speech. *J. Neurosci.* **39**, 7564–7575 (2019).
23. H. Weissbart, K. D. Kandylaki, T. Reichenbach, Cortical tracking of surprisal during continuous speech comprehension. *J. Cogn. Neurosci.* **32**, 155–166 (2020).
24. O. Etard, T. Reichenbach, Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *J. Neurosci.* **39**, 5750–5759 (2019).
25. A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, A. L. Giraud, Speech encoding by coupled cortical theta and gamma oscillations. *eLife* **4**, e06213 (2015).
26. K. B. Doelling, L. H. Arnal, O. Ghitza, D. Poeppel, Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* **85**, 761–768 (2014).
27. M. Keshavarzi, M. Kegler, S. Kadir, T. Reichenbach, Transcranial alternating current stimulation in the theta band but not in the delta band modulates the comprehension of naturalistic speech in noise. *Neuroimage* **210**, 116557 (2020).
28. P. Lakatos *et al.*, The leading sense: Supramodal control of neurophysiological context by attention. *Neuron* **64**, 419–430 (2009).
29. C. Kayser, Phase resetting as a mechanism for supramodal attentional control. *Neuron* **64**, 300–302 (2009).
30. C. Kayser, C. I. Petkov, M. Augath, N. K. Logothetis, Integration of touch and sound in auditory cortex. *Neuron* **48**, 373–384 (2005).
31. S. Soto-Faraco, G. Deco, Multisensory contributions to the perception of vibrotactile events. *Behav. Brain Res.* **196**, 145–154 (2009).
32. K. M. G. Fu *et al.*, Auditory cortical neurons respond to somatosensory stimulation. *J. Neurosci.* **23**, 7510–7515 (2003).
33. G. Caetano, V. Jousmäki, Evidence of vibrotactile input to human auditory cortex. *Neuroimage* **29**, 15–28 (2006).
34. P. Lakatos, C. M. Chen, M. N. O'Connell, A. Mills, C. E. Schroeder, Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* **53**, 279–292 (2007).
35. J. J. Foxe *et al.*, Auditory-somatosensory multisensory processing in auditory association cortex: An fMRI study. *J. Neurophysiol.* **88**, 540–543 (2002).
36. J. J. Foxe *et al.*, Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Brain Res. Cogn. Brain Res.* **10**, 77–83 (2000).
37. M. M. Murray *et al.*, Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cereb. Cortex* **15**, 963–974 (2005).
38. C. E. Schroeder *et al.*, Somatosensory input to auditory association cortex in the macaque monkey. *J. Neurophysiol.* **85**, 1322–1327 (2001).
39. T. A. Hackett *et al.*, Multisensory convergence in auditory cortex, II. Thalamocortical connections of the caudal superior temporal plane. *J. Comp. Neurol.* **502**, 924–952 (2007).
40. L. Riecke, S. Snipes, S. van Bree, A. Kaas, L. Hausfeld, Audio-tactile enhancement of cortical speech-envelope tracking. *Neuroimage* **202**, 116134 (2019).
41. M. J. Crosse, G. M. Di Liberto, A. Bednar, E. C. Lalor, The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016).
42. P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL, 2007).
43. P. C. Loizou, G. Kim, Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Trans. Audio Speech Lang. Process.* **19**, 47–56 (2011).
44. D. Beck, N. Le Goff, Contemporary hearing aid amplification: Issues and outcomes in 2018. *J. Otolaryngol. Res* **10**, 00303 (2018).
45. B. E. Walden, R. K. Surr, M. T. Cord, B. Edwards, L. Olson, Comparison of benefits provided by different hearing aid technologies. *J. Am. Acad. Audiol.* **11**, 540–560 (2000).
46. C. Meyer, L. Hickson, What factors influence help-seeking for hearing impairment and hearing aid adoption in older adults? *Int. J. Audiol.* **51**, 66–74 (2012).
47. Z. Chen, Y. Luo, N. Mesgarani, "Deep attractor network for single-microphone speaker separation" in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, Piscataway, NJ, 2017), 246–250.
48. K. Cieśla *et al.*, Immediate improvement of speech-in-noise perception through multisensory stimulation via an auditory to tactile sensory substitution. *Restor. Neurol. Neurosci.* **37**, 155–166 (2019).
49. B. Gick, K. M. Jóhannsdóttir, D. Gibraiel, J. Mühlbauer, Tactile enhancement of auditory and visual speech perception in untrained perceivers. *J. Acoust. Soc. Am.* **123**, EL72–EL76 (2008).
50. K. L. Galvin, P. J. Blamey, R. S. Cowan, M. Oerlemans, G. M. Clark, Generalization of tactile perceptual skills to new context following tactile-alone word recognition training with the Tickle Talker. *J. Acoust. Soc. Am.* **108**, 2969–2979 (2000).
51. J. E. Peelle, M. H. Davis, Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* **3**, 320 (2012).
52. C. F. Sambo *et al.*, Visual and spatial modulation of tactile extinction: Behavioural and electrophysiological evidence. *Front. Hum. Neurosci.* **6**, 217 (2012).
53. S. T. Yang, S. H. Jin, G. Lee, S. Y. Jeong, J. An, "Dominant and subdominant hand exhibit different cortical activation patterns during tactile stimulation: An FNIRS study" in *2018 6th International Conference on Brain-Computer Interface (BCI)* (IEEE, Piscataway, NJ, 2018), pp. 1–3.
54. R. A. Stevenson *et al.*, Identifying and quantifying multisensory integration: A tutorial review. *Brain Topogr.* **27**, 707–730 (2014).
55. J. E. Peelle, The hemispheric lateralization of speech processing depends on what "speech" is: A hierarchical perspective. *Front. Hum. Neurosci.* **6**, 309 (2012).
56. S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, Inc., Newton, MA, 2009).
57. C. Benoît, M. Grice, V. Hazan, The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Commun.* **18**, 381–392 (1996).
58. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi" in *Interspeech* (2017), vol. **2017**, pp. 498–502.
59. J. P. Goldman, "Easyalign: An automatic phonetic alignment tool under Praat" in *Interspeech'11: 12th Annual Conference of the International Speech Communication Association* (2011), pp. 3233–3236.
60. A. Gramfort *et al.*, MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, 267 (2013).
61. A. de Cheveigné, D. Arzounian, Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *Neuroimage* **172**, 903–912 (2018).
62. I. Winkler, S. Haufe, M. Tangermann, Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behav. Brain Funct.* **7**, 1–15 (2011).
63. T. Amemiya, Non-linear regression models. *Handb. Econom* **1**, 333–389 (1983).
64. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B (Methodological)* **57**, 289–300 (1995).
65. R. F. Barber *et al.*, Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).
66. P. Guilleminot, T. Reichenbach, Audio-tactile speech. Audio-tactile speech EEG dataset. https://zenodo.org/record/5512578. Deposited 16 September 2021.