# Distracting Linguistic Information Impairs Neural Entrainment to Attended Speech

Bohan Dai[1, 2, *], James M. McQueen[1, 2], René Terporten[1, 2], Peter Hagoort[1, 2], Anne Kösem[1-3]

[1]Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands

[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6500 HB Nijmegen, The Netherlands

[3] Lyon Neuroscience Research Center (CRNL), Brain Dynamics and Cognition Team, INSERM U1028, CNRS UMR5292, Université Claude Bernard Lyon 1, UdL, Lyon, France

***Corresponding Author:** Bohan Dai, Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands. Email: bohan.dai@mpi.nl

## Abstract

Listening to speech is difficult in noisy environments, and is even harder when the interfering noise consists of intelligible speech as compared to non-intelligible sounds. This suggests that the ignored speech is not fully ignored, and that competing linguistic information interferes with the neural processing of target speech. We tested this hypothesis using magnetoencephalography (MEG) while participants listened to target clear speech in the presence of distracting noise-vocoded signals. Crucially, the noise vocoded distractors were initially unintelligible but were perceived as intelligible speech after a small training session. We compared participants' performance in the speech-in-noise task before and after training, and neural entrainment to both target and distracting speech. The comprehension of the target clear speech was reduced in the presence of intelligible distractors as compared to when they were unintelligible. The neural entrainment to target speech in the delta range (1–4 Hz) reduced in strength in the presence of an intelligible distractor. In contrast, neural entrainment to distracting signals was not significantly modulated by intelligibility. These results support and extend previous findings, showing, first, that the masking effects of distracting speech originate from the degradation of the linguistic representation of target speech, and second, that delta entrainment reflects linguistic processing of speech.

## Keywords

Cocktail party, informational masking, neural oscillations, delta, MEG

## Significance Statement

Comprehension of speech in noisy environments is impaired due to interference from background sounds. The magnitude of interference depends on the intelligibility of the distracting speech signals. In a magnetoencephalography experiment with a highly-controlled training paradigm, we show that the linguistic information of distracting speech imposes higher-order interference on the processing of the target speech, as indexed by a decline of comprehension of target speech and a reduction of delta entrainment to target speech. This work demonstrates the importance of neural oscillations for speech processing. It shows that delta oscillations reflect linguistic analysis during speech comprehension, which can critically be affected by the presence of other speech.

## Introduction

Speech communication in everyday life often takes place in the presence of multiple talkers or background noise. In such complex auditory scenes, comprehension of target speech can be degraded due to interference from concurrent sounds (1, 2). Given that speech is a complex auditory signal that carries linguistic information, the competition between target speech and distracting sounds occurs at multiple levels of the speech processing hierarchy: interference could occur during the auditory analysis of speech, or at a later stage during the decoding of linguistic information (3–6). Interfering signals with different types of acoustic and/or linguistic information should thus influence different aspects of the neural processing of target speech. In line with this prediction, the comprehension of target speech is known to depend on the acoustic complexity of the distracting sounds (7, 8). On top of acoustic effects, the intelligibility of distracting speech alone is also a source of interference, such that the same noise-vocoded speech background impairs more strongly the comprehension of target speech when it is intelligible as compared to when it is unintelligible (9).

Behavioral evidence highlights that distracting sounds do not only mask the acoustics of attended signals, but also interfere with the processing of abstract features of target speech. The neural origins of this interference, however, are still unclear. Interference could either arise from a degradation of the neural representation of the target speech, or from increased representation of distracting speech that enters in competition with the target speech. To test these alternative hypotheses, we used selective neural entrainment to speech dynamics as a measure of brain-speech alignment. Degree of entrainment indicates how well the attended speech stream is segregated from the listening background (10–12). When listening to clear speech, neural oscillatory activity in the delta (1–4 Hz) and theta (4–8 Hz) ranges entrain to the dynamics of speech (10, 13–18). In a noisy or multi-talker scene, both theta and delta neural oscillations primarily entrain to the dynamics of the attended speech (10, 12, 14, 19). Yet, it is still under debate which aspects of speech are encoded in entrainment activity. Broadly speaking, this could be either acoustic features or higher-level language information (20, 21). Recent work suggests that the different neural oscillatory markers link to distinct aspects of speech perception: theta entrainment underlies speech sound analysis (22–24) while delta oscillations reflect higher-level linguistic processing, such as semantic and syntactic processing (15, 25).

Here, we examined whether competing linguistic information influences the neural entrainment to target and ignored speech in a cocktail-party setting. Based on previous findings, we hypothesized that linguistic masking should impact delta entrainment. In order to isolate the linguistic effect from the effect of acoustic competition between the target and distracting speech, we used a novel A-B-A training paradigm in which the linguistic content of the distracting stimulus was manipulated while its acoustic properties were kept constant (9). Participants performed a dichotic listening task twice, in which they were asked to repeat a clear speech signal while noise-vocoded (NV) speech was presented as distractor (Fig. 1A). In between the two sessions, participants were trained to understand the interfering NV distractor (Fig. 1B). We compared behavioral performance (accuracy in the repetition of the target speech) and MEG oscillatory activity between the two dichotic listening sessions. Our main prediction was that distracting speech would impair more strongly target speech comprehension when intelligible, and that the linguistic masking would modulate the pattern of neural entrainment to target speech.

--- Insert Figure 1 here---

## Results

### *Intelligible NV speech interfered more with target speech's understanding*

Two types of NV speech were used as distractors in the dichotic listening task: either 4-band or 2-band NV speech segments. In the training phase, participants were trained to understand 4-band NV speech, while 2-band NV speech was not trained. Hence, 2-band NV speech served as control distractors that would not improve in intelligibility with training. To make sure the training was efficient in improving the intelligibility of 4-band NV speech, we compared the participants' comprehension of the NV signals before and after training. Consistent with previous findings (9), and as shown in Fig. 2A, the training significantly improved the perception of 4-band NV speech. A two-way repeated-measure ANOVA showed that the main effects of noise vocoding (2-band *vs*. 4-band) and time (pre- *vs*. post-training) were significant (noise vocoding: ($F(1, 24) = 217.78$, $p < 0.001$; time: $F(1, 24) = 219.07$, $p < 0.001$). Crucially, a significant interaction between noise vocoding and time was observed ($F(1,24) = 262.94$, $p < 0.001$), meaning that the intelligibility of 4-band NV speech was significantly improved compared to that of 2-band NV speech (4-band(post-pre) vs. 2-band(post-pre): $t$ (24) = 16.22, $p < 0.001$). After training, 4-band NV sentences had a score of 52.69 ± 3.35% recognition accuracy (31.70 ± 2.03% improvement during training; values here and below indicate mean ± SEM), while 2-band NV sentences remained mostly unintelligible with a score of 1.97 ± 0.52% recognition accuracy (1.06 ± 0.35% improvement during training).

--- Insert Figure 2 here---

The training efficiently improved the intelligibility of 4-band NV speech. We then investigated if the change in the intelligibility of the distractor interfered with the comprehension of the target speech during dichotic listening. To assess the magnitude of increased interference, we measured the accuracy of target speech recognition in the two dichotic listening tasks (Fig. 2B). A three-way repeated-measure ANOVA was performed (time (pre-training, post-training), noise vocoding (trained 4-band, untrained 2-band), and side of target presentation (left target, right target)). We observed a significant main effect of noise vocoding ($F(1, 24) = 6.44$, $p < 0.05$), and a significant interaction of the three factors ($F(1, 24) = 8.22$, $p < 0.01$), which revealed that the change of interference depended on which ear the target speech was delivered. A closer look at the data showed that target speech comprehension decreased after training when the target speech was presented to the left ear (i.e., when the distractor NV signal was presented to the right ear). There was a significant interaction between noise vocoding and time ($F(1, 24) = 8.12$, $p < 0.01$): the 4-band NV speech interfered more strongly with target speech comprehension after training than before training (pre = 97.22 ± 0.42%; post = 94.42 ± 1.11%; post vs pre: $t$ (24) = -3.08, $p < 0.01$), while 2-band NV speech had similar masking effect before and after training (pre = 96.54± 0.61%; post = 96.91 ± 0.49%; post vs pre: $t$ (24) = - 0.93, $p$ = 0.362). As previously shown (9), these findings suggest that the increased intelligibility of 4-band NV speech acquired during training generates more interference in the processing of the target speech and decreases its comprehension. When the target speech was delivered to the right ear (and the distractor to the left ear), no effect of training was observed (interaction of noise-vocoding and time: $F(1, 24) = 0.003$, $p = 0.954$; 4-band: pre = 96.30 ± 0.54%, post = 95.80 ± 1.05; 2-band: pre = 96.85 ± 0.46%, post = 96.29 ± 0.65%). This is in line with previous studies showing a right ear advantage in speech processing (14, 26–29). In our study, when the distractor was displayed on the right ear, it is primarily processed in the left language-dominant hemisphere, and may have its processing facilitated thus offering stronger interference when it is more intelligible (12). Effects of ear of presentation were not reported in Dai et al. (2017), but additional analysis of those data revealed a similar (but not statistically significant) asymmetric pattern.

--- Insert Figure 3 here---

***Neural entrainment to target speech and distracting speech***

In the first stage of MEG analysis, we inspected the speech-brain coherence for both target speech and distracting speech (Fig. 3). We focused on both delta (1–4 Hz) and theta (4–8 Hz) entrainment to speech, as both frequency ranges are deemed relevant for speech processing (10, 14–18, 21, 30, 31). Coherence data were computed from the 36 channels (18 channels in each hemisphere) that produced the strongest auditory evoked M100 responses (Fig. 3), and were first averaged across all conditions (i.e., across the pre- and post-training sessions and across distractor type) and all sensors. A three-way repeated-measure ANOVA was performed (frequency (delta, theta), speech type (target, distractor), and data (data, surrogate)). Compared to surrogate coherence between neural oscillations and speech envelope, we observed stronger target- and distractor-brain coherence in both delta and theta ranges (main effect of data: ($F(1, 24) = 61.86$, $p < 0.0001$; Fig. 3A-B). A main effect of frequency was observed as well ($F(1, 24) = 142.874$, $p < 0.0001$), showing stronger speech-brain coherence in the delta range than in the theta range (Fig. 3A-B). Overall, entrainment to target speech was stronger than that to distracting speech (main effect of speech type: ($F(1, 24) = 62.15$, $p < 0.0001$, interaction of data and speech type ($F(1, 24) = 63.72$, $p < 0.0001$, post-hoc tests: data, target > distractor, $p < 0.0001$; surrogate, target vs. distractor, $p = 0.176$, Bonferroni corrected), and this specifically for delta oscillations (interaction between frequency and speech type ($F(1, 24) = 8.39$, $p < 0.01$, post-hoc tests: delta, target > distractor, $p < 0.0001$; theta, target > distractor, $p < 0.0001$, Bonferroni corrected, Fig. 3A-B). As we used a dichotic listening task, speech-brain coherence was stronger on the contralateral side to the ear of presentation, for both target and distracting speech (Fig 3C-D). All the interactions of side and hemisphere were significant: delta entrainment to target speech: $F(1, 24) = 14.79$, $p < 0.001$; delta entrainment to distracting speech: $F(1, 24) = 7.91$, $p < 0.05$; theta entrainment to target speech: $F(1, 24) = 36.48$, $p < 0.001$; theta entrainment to distracting speech: $F(1, 24) = 21.64$, $p < 0.001$). However, speech-brain coherence for both target and distracting speech were also observed on the ipsilateral side, suggesting that both signals were processed bilaterally even in a dichotic listening task.

### *Neural entrainment to target speech was modulated by linguistic information of the distracting speech*

We then tested the effect of training on speech-brain coherence in the delta and theta ranges. Specifically, we expected the neural analysis of target speech (as reflected by neural entrainment) to be more impaired in the presence of an intelligible distractor. Hence, we predicted speech-brain coherence to target speech to become weaker after training, and this only for the 4-band NV distractor condition as the effect of training was limited to this type of distracting signal.

--- Insert Figure 4 here---

This prediction was supported by target-brain coherence: delta entrainment to target speech was reduced after training when distractor signals were 4-band NV speech, while the 2-band NV speech condition did not show this change (Fig. 4, interaction between noise-vocoding and time ($F(1, 24) = 5.21$, $p = 0.032$), post-hoc effects: 4-band, post vs. pre, $p < 0.001$; 2-band, post vs. pre: $p = 0.60$, Bonferroni corrected for multiple comparisons). To test whether the behavioral reduction derived from the neural changes, we correlated the relative changes of entrainment between pre- and post-training sessions (entrainment change = $(\text{Coh}_{\text{post}} - \text{Coh}_{\text{pre}}) / (\text{Coh}_{\text{post}} + \text{Coh}_{\text{pre}})$) with the absolute behavioral change. A significant correlation between the relative change of delta entrainment to target speech in the left hemisphere and the behavioral reduction of reporting target speech was observed, but only when target speech was delivered to the left ear (Fig. 5, $rho = 0.53$, $p = 0.028$, Bonferroni corrected for multiple comparisons).

--- Insert Figure 5 here---

In contrast to delta oscillations, target-brain coherence in the theta range was not significantly affected by the intelligibility of the distractor (Fig. 1S). Theta entrainment to target speech did show a significant interaction of the factors time, noise-vocoding and side of presentation. However, the post-hoc tests examining the interaction yielded no significant difference.

### *Neural entrainment to distracting speech was not modulated by linguistic information of the distracting speech*

We also tested whether the increased intelligibility of the distractor had an effect on the distractor-brain coherence. As previous studies suggested that neural entrainment is stronger for intelligible signals (32, though see 24, 33), we asked whether speech-brain coherence to distracting signals would increase after training for the intelligible 4-band NV distractor sentences. However, this effect was not present in the data. Distractor-brain coherence in the delta frequency was overall reduced after training compared to before training (Fig. 6, main effect of time: $F(1, 24) = 37.85$, $p < 0.0001$) irrespective of the type of distractor (2-band or 4-band NV speech). The reduction of delta entrainment to distracting speech can thus not be attributed to the training or the degree of intelligibility of the distractor, but may relate to habituation effects. Similarly, theta entrainment of distracting speech was not modulated via training in our experiment (Fig. 2S). These results suggest that distractor speech-brain coherence is not influenced by the linguistic properties of the distracting signal.

--- Insert Figure 6 here---

## Discussion

In this MEG study, we developed a new training paradigm with which we were able to separate the linguistic and acoustic components of the masking effect between two speech signals. Our data show that distracting speech can exert stronger interference on the processing of target speech when it becomes more intelligible. This increased interference reduced the neural tracking of target speech in the brain. Altered entrainment to target speech could represent a crucial influence on its comprehension when it is heard together with intelligible distractor speech. Moreover, neural oscillations at multiple time scales likely played different roles during speech processing: the neural entrainment to target speech reduced in the delta range (1–4 Hz) in the presence of an intelligible distractor but did not do so in the theta range (4–8Hz). Overall, our results suggest a hierarchy of masking effects in auditory scene analysis.

Since the classic work on the cocktail-party problem 60 years ago (2), researchers have put a lot of effort into understanding the competing processing of target speech and background signals (7–10, 12, 19, 34, 35). It has been suggested that distracting signals exert influence on understanding target speech depending on the amount of linguistic information. For example, researchers have shown that speech signals impair comprehension more strongly than unintelligible sounds (7, 8). However, the previous studies often manipulated the intelligibility of distracting sounds that typically affect both acoustic and linguistic content (e. g. speech *vs.* reversed speech, or native *vs.* non-native speech), leaving distinctions between acoustic interference and linguistic interference unresolved. We used a training paradigm (9) which allowed us to manipulate the intelligibility of distracting speech without changing its acoustic component, and therefore isolated the higher-order linguistic competition from lower-order effects. Our results demonstrate that intelligible speech is a stronger distractor than unintelligible speech. Given the training manipulation, stronger masking is not due to the similarity of acoustic aspects between the target and the distracting speech. Instead, it reflects effects of the higher-order linguistic information that can be extracted from the distracting signals after training. These results certainly do not exclude the possibility that acoustic information in distracting speech can have a masking effect. Rather, they suggest that acoustic masking is only part of the story, and linguistic information offers extra interference.

Our results show that the neural entrainment to target speech in the delta range reduced with a more intelligible distracting speech; while entrainment in the theta range did not change with intelligibility. This is in line with main neural frameworks of speech analysis, suggesting that higher-level linguistic processing often involves neural oscillations with longer time scales compared to lower-level analysis (17, 21, 36). Specifically, neural entrainment in the delta range has been linked to the encoding of linguistic information (15, 24, 25), while theta oscillations may primarily relate to acoustic analysis (21). In multi-speech scenes, both the target and distracting speech have multi-level information ranging from their acoustic features to linguistic meanings, and therefore their competition could happen on each level of the hierarchy. With this dichotic listening task, we thus demonstrate a hierarchical system of competition between the two signals.

We did not observe a significant change in the neural entrainment to distracting signals with intelligibility. The link between strength of entrainment and intelligibility is debated due to the contradictory findings: studies have reported that low-frequency neural entrainment is stronger when speech is intelligible (13, 32, 37–39), while other studies failed to find a correlation between neural entrainment to speech and intelligibility (24, 33, 40, 41). A likely source of the different results is acoustic confounds (21). Here, we carefully controlled for acoustic confounds and did not find significant evidence that neural entrainment to distracting signals increased in strength with its intelligibility. However, we found that distractor intelligibility decreased the entrainment strength to target speech. This suggest that the masking effect (the increased misunderstanding of the target speech) in our task originated from the degradation of the neural representation of the target speech, and not from the increased neural representation of the competing speech.

Despite presenting the target and distracting signals in different ears, we showed that both distracter and target speech signals were processed to some extent in the ipsilateral auditory cortex. Furthermore, the effect of distracter intelligibility on neural entrainment to target was observed in both hemispheres and irrespective of the ear of presentation of target and distracter speech. However, we observed an effect of the ear of presentation on behavioral performance in line with previous reports (14, 26–29). An intelligible NV distracter impaired more strongly target speech comprehension when it was presented to the right ear and the target was presented to the left ear, that is, when the distracter was primarily processed in the left hemisphere. This effect could be explained by the fact that, in this scenario, distractor signals have prior access to the language processing network which is known to be left lateralized. The processing of distracting linguistic information is facilitated, and this could cause stronger interference on the linguistic processing of target speech. In line with this interpretation, the neural entrainment to target speech in the left hemisphere was associated with loss in target intelligibility.

In summary, our data provide evidence that, in a multi-talker environment, the linguistic information of distracting speech imposes higher-order interference on the processing of the target speech, as indexed by a reduction of delta entrainment to target speech. The decrease in target speech entrainment is correlated with the decline of target speech comprehension. The findings from this highly-controlled training paradigm show that delta oscillations reflect speech-specific analysis during comprehension of spoken language.

## Methods

### *Participants*

Twenty-seven participants (13 women, mean age: 23.5 ± 3.9 years) took part in the study. All were right-handed native Dutch speakers with normal hearing. Two participants were rejected, one due to malfunctioning of the MEG system and one because the participant did not finish the task. The analyses are thus based on the data of 25 participants (12 women, mean age: 23.5 ± 4.0 years). All participants gave their informed consent in accordance with the Declaration of Helsinki, and the local ethics committee (CMO region Arnhem-Nijmegen).

### *Stimuli*

As in our previous study (9), the stimuli were selected from a corpus with meaningful conversational Dutch sentences (e.g., 'Mijn handen en voeten zijn ijskoud', in English: 'My hands and feet are freezing'), digitized at a 44,100 Hz sampling rate and recorded at the VU University Amsterdam (42) by a male or female speaker. Each speech stimulus consisted of a combination of two sentences of the corpus uttered by the same speaker, separated by a 300-ms silence gap (average duration = 4.15 ± 0.13 s).

The target speech stimuli consisted of 384 intact sentence pairs spoken by one of the two speakers (half of the trials were from the male speaker and half were from the female speaker). The distracting speech stimuli were NV versions of 48 different sentence pairs taken from the same corpus and spoken by the other speaker (i.e., a speaker of opposite sex). Noise-vocoding (43) was performed using either 4 (main condition) or 2 (control) frequency bands logarithmically spaced between 50 and 8000 Hz. In essence, the noise-vocoding technique parametrically degrades the spectral content of the acoustic signal (i.e., the fine structure) but keep the temporal information largely intact.

### *Procedure*

The main experiment was similar to our previous study (9). The experimental design was implemented using Presentation software (Version 16.2, www.neurobs.com).

The experiment included three phases: pre-training, training, and post-training. In the pre- and post-training phases, the participants performed the dichotic listening task. Each trial consisted of the presentation of the target speech with the interfering NV speech. The two signals were delivered dichotically to the two ears. The target side (left or right) for a particular trial was pseudo-randomly defined: half of the trials had the target on the left ear and half had it on the right ear. The stimuli were presented at a comfortable listening level (70 dB) by MEG-compatible air tubes in a magnetically shielded room. The signal-to-noise ratio (SNR) was fixed on -3 dB based on the results of our previous study, -3dB SNR being the condition in which we observed the strongest masking effects of intelligible distractor signals. The participants were instructed to listen to the presentation of one intact speech channel and one unintelligible NV speech channel and pay attention to the intact target speech only. After the presentation, the participant's task was to repeat the sentences of the target speech. Participants' responses were recorded by a digital microphone with a sampling rate of 44,100 Hz. The distracting speech consisted of 24 sentences of 4-band NV speech and 24 sentences of 2-band NV speech.  Each dichotic listening task comprised 192 trials total. The target speech differed across trials and all conditions (hence a target stimulus was only presented once during the whole experiment), while NV distracting stimuli were repeated four times, for a total of 96 trials per NV condition. The ear of presentation of the target signals, and the type of NV signals (4-band, 2-band), were randomized across trials. Each dichotic listening task was 30 min long.

In the training phase, participants were trained to understand the 4-band NV speech. The training phase included three parts: (a) pre-test: the participants were tested on their ability to understand the 24 4-band NV stimuli and 24 2-band NV stimuli used in the dichotic listening task as distracting signals; they were presented with the interfering speech binaurally and were asked to repeat it afterwards; (b) training on 4-band NV speech: they were presented one token of an intact version of an NV stimulus followed by one token of the NV version of that stimulus; at the same time, they could read the content of the NV speech on the screen; 2-band NV speech was not trained; (c) post-test: they performed the intelligibility test again. Crucially, the 4-band NV speech were initially poorly intelligible but could be understood after training (9, 43, 44). Hence, during the pre- and post-training phases, the NV speech would have the same acoustic information but would not allow for extraction of the same amount of linguistic information. In total, the training phase took 30 min.

### *Behavioral analysis*

The intelligibility of speech was measured by calculating the percentage of correct content words (excluding function words) in participants' reports for each trial. Words were regarded as correct if there was a perfect match (correct word without any tense errors, singular/plural form changes, or changes in sentential position). The percentage of correct content words was chosen as a more accurate measure of intelligibility based on acoustic cues than percentage correct of all words, considering that function words can be guessed based on the content words (45). For the training phase, we performed a two-way repeated-measures ANOVA with noise vocoding (trained 4-

band and untrained 2-band) and time (pre- and post-training) as factors. For the dichotic listening tasks, a three-way repeated-measures ANOVA was performed to assess the contribution of three factors: noise vocoding (4-band and 2-band), time (pre-training and post-training), and side (left target/right distractor and right target/left distractor). In our post hoc sample t-tests, we compensated for multiple comparisons with Bonferroni correction.

### *MEG Data Acquisition*

MEG data were recorded with a 275-channel whole-head system (CTF Systems Inc., Port Coquitlam, Canada) at a sampling rate of 1200 Hz in a magnetically shielded room. Data of two channels (MLC11 and MRF66) were not recorded due to channel malfunctioning. Participants were seated in an upright position. Head location was measured with two coils in the ears (fixed to anatomical landmarks) and one on the nasion. To reduce head motion, a neck brace was used to stabilize the head. Head motion was monitored online throughout the experiment with a real-time head localizer and if necessary corrected between the experimental blocks.

### *MEG Data preprocessing*

Data were analyzed with the FieldTrip toolbox implemented in MATLAB (46). Trials were defined as data between 500 ms before the onset of sound signal and 4, 000 ms thereafter. Three steps were taken to remove artifacts. Firstly, trials were rejected if the range and variance of the MEG signal differed by at least an order of magnitude from the other trials of the same participant. On average, 14.1 trials per participant were rejected (SD = 7.8) via visual inspection. Secondly, data were down-sampled to 100 Hz and independent component analysis (ICA) was performed. Based on visual inspection of the ICA components' time courses and scalp topographies, components showing clear signature of eye blinks, eye movement, heartbeat and noise were identified and removed from the data. On average, per participant 9.8 components (SD = 2.6) were rejected (but no complete trials). Finally, 9.6 trials (SD = 4.4) with other artifacts were removed via visual inspection like the first step, resulting in an average of 360 trials per participant (each condition: ~90 trials). Subsequently, the clean data were used for further analyses.

### *MEG analysis*

A data-driven approach was performed to identify the reactive channels for sound processing. The M100 (within the time window between 80ms and 120ms after the first world were presented) response was measured on the data over all experimental conditions, after planar gradient transformation (47).We selected the 18 channels with the relatively strongest response on each hemisphere, and the averages of these channels were used for all subsequent analysis. The locations of the identified channels cover the classic auditory areas.

Speech-brain coherence analysis was performed on the data within the region of interest after planar gradient projection. Spectral analysis was performed using 'dpss' multi- tapers with a ± 1 Hz smoothing window of the speech envelopes, and of the neural times series epoched from 500 epochs were removed to exclude the evoked response to the onset of the sentence. To match trials in duration, shorter trials were zero-padded up to 3.4s (the max length of the signal) for both target speech and distracting speech. For the plots in Figure 1, the speech-brain coherence was measured at different frequencies (1 to 15 Hz, 1 Hz step). Finally, the coherence data were projected into planar gradient representations. Then data were averaged across all trials and the strongest 36 channels defined by our auditory response localizer. Following the same method, we calculated the surrogate of speech-brain coherence (as control condition) by randomly selecting the neural activity of one trial and the temporal envelope of speech of another trial. For the investigation of our main hypotheses, we restricted the speech-brain coherence analyses to delta band (1–4 Hz) and theta band (4–8 Hz) activity; these frequency bands were chosen based on the previous literature (10, 14, 16, 17, 21). A three-way repeated-measure ANOVA was performed (frequency (delta, theta), speech type (target, distractor), and data (data, surrogate)). We repeated the same analysis described above to quantify the speech-brain coherence for each condition, and the averaged speech-brain coherences of the strongest 18 channels on each hemisphere in the delta and theta bands were calculated. We tested the target and distractor speech-brain coherence in the delta and theta range using a four-way repeated measure ANOVA with factors noise vocoding (4-band, 2-band), time (pre-training, post-training), side (left target, right target), and hemisphere (left, right). The relative coherence change of training was calculated based on the following formula: relative change = $(\mathrm{Coh}_{\mathrm{post}} - \mathrm{Coh}_{\mathrm{pre}}) / (\mathrm{Coh}_{\mathrm{post}} + \mathrm{Coh}_{\mathrm{pre}})$. Afterward, the correlation of coherence change and behavioral change was calculated.

### Materials and Data availability

Data and code related to this paper are available upon request from the Donders Repository (https://webdav.data.donders.ru.nl:443/dccn/DAC_3011087.01_349/), a data archive hosted by the Donders Institute for Brain, Cognition and Behaviour.

## Acknowledgments

We would like to thank Ole Jensen for providing insightful suggestions on this work; Maarten van den Heuvel and Vera van 't Hoff for help with transcribing data. This research was supported by a Spinoza grant to P.H.

## Author contributions

B.D., A.K., and P.H. conceived and designed research; B.D. performed research; B.D. and A.K. analyzed the data; all authors contributed to interpretation of the results; B.D., A.K., J.M.M, and P. H. wrote the paper.

## Conflict of interests

The authors declare no conflict of interests.

## References

1. Carlile S (2015) Auditory perception: attentive solution to the Cocktail Party Problem. *Curr Biol* 25(17):R757–R759.

2. Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 25:975–979.

3. Evans S, Davis MH (2015) Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cereb Cortex* 25(12):4772–4788.

4. Hoen M, et al. (2007) Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech Commun* 49(12):905–916.

5. Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: A review. *Lang Cogn Process* 27(7–8):953–978.

6. Rhebergen KS, Versfeld NJ, Dreschler WA (2005) Release from informational masking by time reversal of native and non-native interfering speech. *J Acoust Soc Am* 118(3):1274–1277.

7. Brungart DS, Simpson BD, Ericson MA, Scott KR (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J Acoust Soc Am* 110:2527–38.

8. Scott SK, Rosen S, Wickham L, Wise RJS (2004) A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *J Acoust Soc Am* 115(2):813.

9. Dai B, McQueen JM, Hagoort P, Kösem A (2017) Pure linguistic interference during comprehension of competing speech signals. *J Acoust Soc Am* 141(3):EL249-EL254.

10. Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci U A* 109:11854–9.

11. Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320(5872):110–113.

12. Zion Golumbic EM, et al. (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron* 77:980–91.

13. Ahissar E, et al. (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci* 98(23):13367–13372.

14. Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107:78–89.

15. Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19(1):158–164.

16. Gross J, et al. (2013) Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol* 11(12):e1001752.

17. Luo H, Poeppel D (2012) Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Front Psychol* 3:170.

18. Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54(6):1001–1010.

19. Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–6.

20. Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8. doi:10.3389/fnhum.2014.00311.

21. Kösem A, Wassenhove V van (2017) Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Lang Cogn Neurosci* 32(5):536–544.

22. Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol* 25(19):2457–2465.

23. Kösem A, Basirat A, Azizi L, Wassenhove V van (2016) High-frequency neural activity predicts word parsing in ambiguous speech streams. *J Neurophysiol* 116(6):2497–2512.

24. Millman RE, Johnson SR, Prendergast G (2014) The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *J Cogn Neurosci* 27(3):533–545.

25. Di Liberto GM, Lalor EC, Millman RE (2018) Causal cortical dynamics of a predictive enhancement of speech intelligibility. *NeuroImage* 166(Supplement C):247–258.

26. Berlin CI, Lowe-Bell SS, Cullen JK, Thompson CL, Loovis CF (1973) Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects. *J Acoust Soc Am* 53(3):699–709.

27. Brancucci A, et al. (2004) Inhibition of auditory cortical responses to ipsilateral stimuli during dichotic listening: evidence from magnetoencephalography. *Eur J Neurosci* 19(8):2329–2336.

28. Della Penna S, et al. (2007) Lateralization of dichotic speech stimuli is based on specific auditory pathway interactions: neuromagnetic evidence. *Cereb Cortex* 17(10):2303–2311.

29. Hiscock M, Kinsbourne M (2011) Attention and the right-ear advantage: What is the connection? *Brain Cogn* 76(2):263–275.

30. Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15(4):511–517.

31. Park H, Ince RAA, Schyns PG, Thut G, Gross J (2015) Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr Biol* 25(12):1649–1653.

32. Peelle JE, Gross J, Davis MH (2012) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23(6):1378–1387.

33. Zoefel B, VanRullen R (2016) EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage* 124:16–23.

34. Scott SK, Rosen S, Beaman CP, Davis JP, Wise RJS (2009) The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. *J Acoust Soc Am* 125(3):1737–1743.

35. Dai B, et al. (2018) Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. *Nat Commun* 9(1):2405.

36. Poeppel D (2014) The neuroanatomic and neurophysiological infrastructure for speech and language. *Curr Opin Neurobiol* 28:142–149.

37. Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J Neurosci* 33:5728–35.

38. Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85, Part 2:761–768.

39. Zoefel B, Archer-Boyd A, Davis MH (2018) Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Curr Biol* 28(3):401–408.e5.

40. Howard MF, Poeppel D (2010) Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol* 104:2500–11.

41. Peña M, Melloni L (2012) Brain oscillations during spoken sentence processing. *J Cogn Neurosci* 24(5):1149–1164.

42. Versfeld NJ, Daalder L, Festen JM, Houtgast T (2000) Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J Acoust Soc Am* 107(3):1671–1684.

43. Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C (2005) Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen* 134(2):222–241.

44. Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. *Proc Natl Acad Sci* 113(12):E1747–E1756.

45. Brouwer S, Van Engen KJ, Calandruccio L, Bradlow AR (2012) Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *J Acoust Soc Am* 131(2):1449–1464.

46. Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Intell Neurosci* 2011:1:1–1:9.

47. Oever S ten, et al. (2017) Low-frequency cortical oscillations entrain to subthreshold rhythmic auditory stimuli. *J Neurosci* 37(19):4903–4912.

## Figure Legends

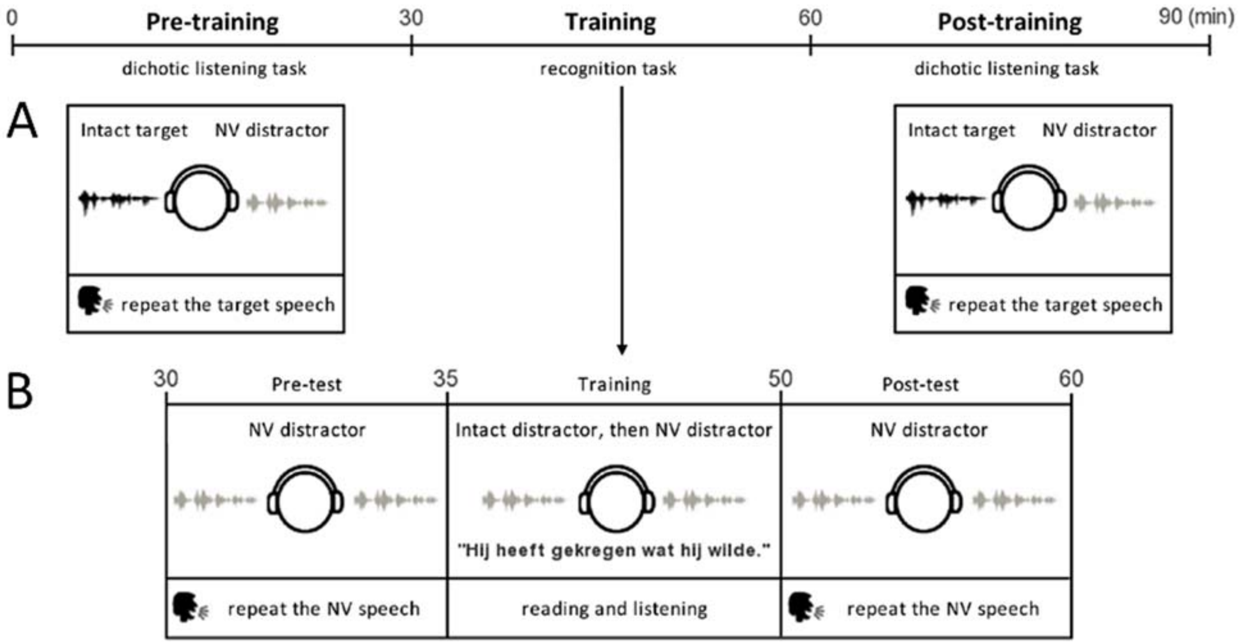

**Figure 1. Experimental design**. The experiment consisted of three phases. In the first and third phases of the experiment, participants performed a dichotic listening task (A), in between the two dichotic listening tasks participants were trained to understand 4-band NV speech (B). (A) During the dichotic listening tasks, participants listened to the presentation of one intact target with one NV distractor (either 2- or 4-band) and were asked to repeat the intact target speech. (B) During the training of the 4-band NV sentences, participants listened to the distractor once in the intact and then once in the NV version. At the same time, they read the text of the sentences on the screen. We tested the intelligibility of the 4-band NV sentences before (pre-test) and after (post-test) the training by asking participants to listen to and repeat the NV sentences.

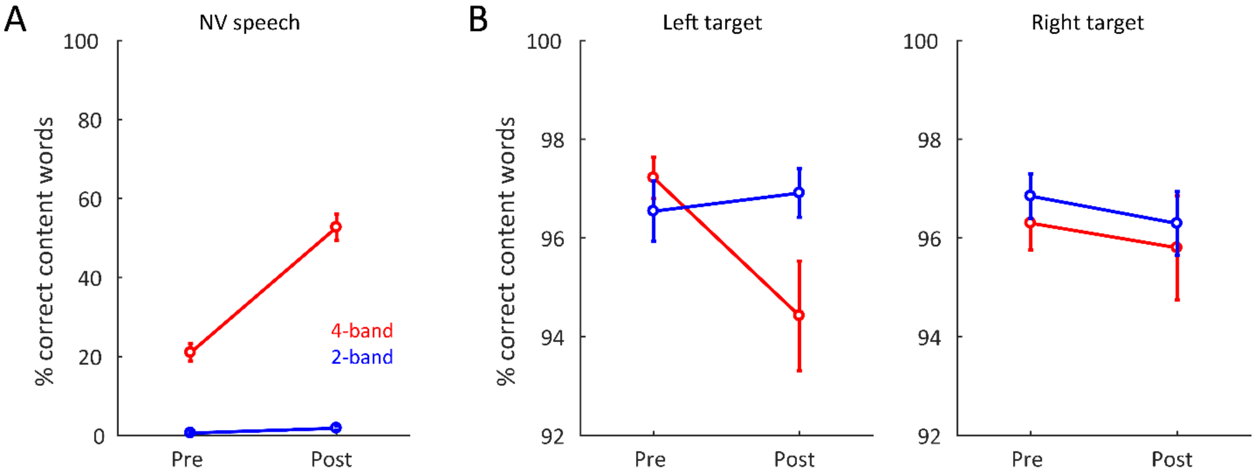

**Figure 2. Behavioral results.** (A) Intelligibility of NV speech during the training phase. The intelligibility of trained 4-band (red) significantly improved by more than 30% with training, while untrained 2-band (blue) NV speech remained mostly unintelligible post-training. (B) Intelligibility of target speech in the dichotic listening tasks. The intelligibility of target speech decreased after training when presented in competition with the 4-band NV speech (red), i.e. when distracting NV speech was more intelligible. The intelligibility of target speech was not significantly affected by training when presented in competition with the untrained 2-band NV speech (blue). This is only observed when the target speech is delivered to the left ear and the distractor to the right ear (left panel) and not when the target is delivered to the right ear and the distractor to the left ear (right panel). Error bars indicate standard error of the mean.

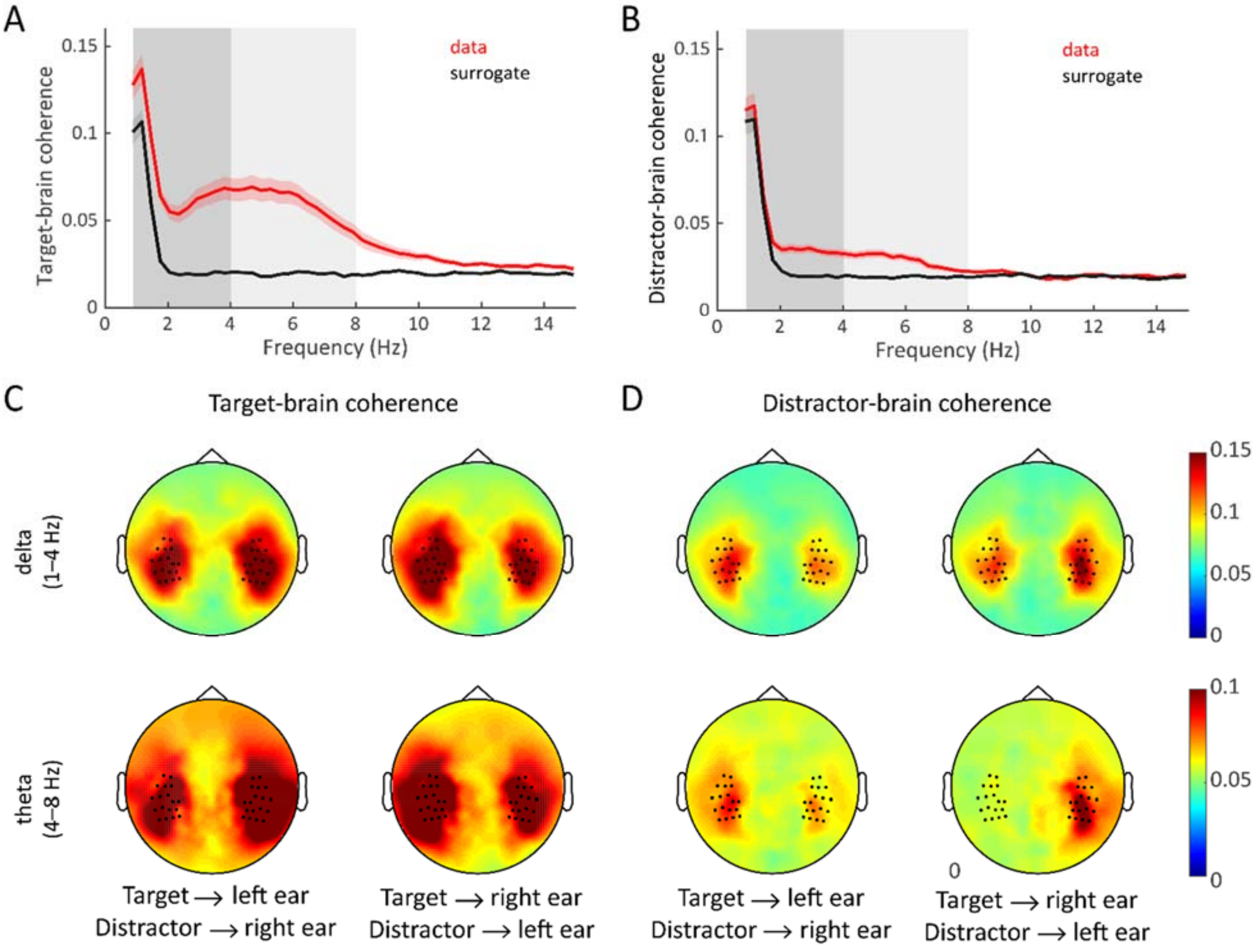

**Figure 3. Neural entrainment to target and distracting speech in both dichotic listening tasks.** The top panel respectively show neural entrainments to (A) target and (B) distracting speech from 1 to 15 Hz. The red line is the real speech-brain coherence data, while the black line is the control data which was calculated by randomly combining brain and speech data from different trials (see methods for details). The dark and light grey shadows mark respectively the delta and theta ranges as defined in our study. (C) and (D) Topographies of delta and theta entrainments to (C) target speech and to (D) distracting speech averaged across all conditions. Speech-brain coherence is stronger in the auditory regions contralateral to the ear of presentation of the stimulus, though brain responses are observed bilaterally. Black dots mark the selected channels for the further speech-brain coherence analyses.

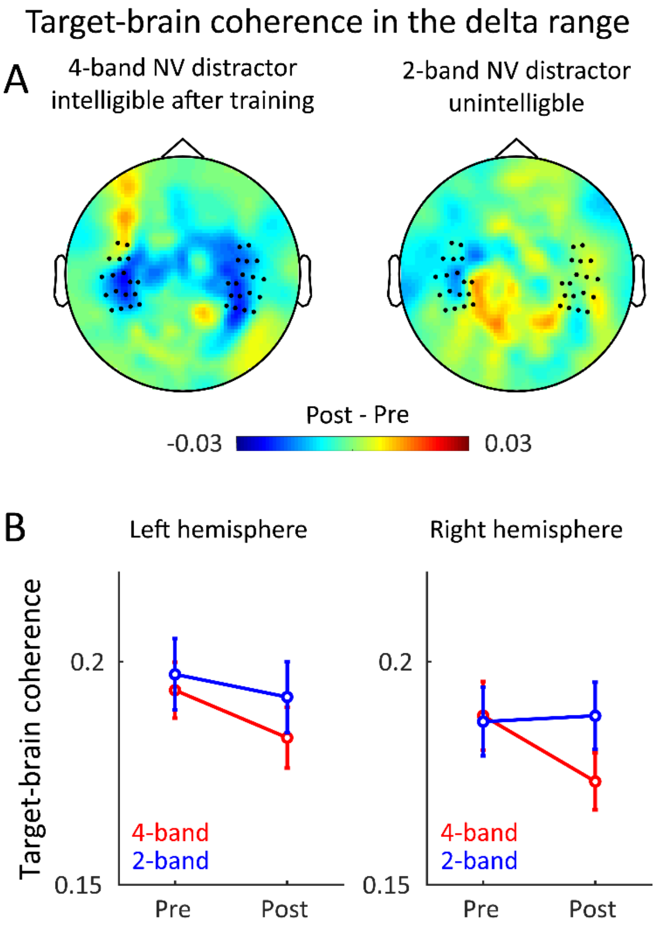

**Figure 4. Delta neural entrainment to target speech reduced when the distracting NV speech gained intelligibility**. (A) Topographies of delta entrainment changes (Post-training minus Pre-training). A significant reduction in delta entrainment to target was observed in bilateral temporal lobes after training when it was presented in competition with a distracting 4-band NV speech, i.e. when the distracting speech had gained intelligibility via training (left panel). No significant effects of training were observed when the target speech was presented with unintelligible 2-band NV distracters (right panel) (B) Delta entrainments to target speech averaged across selected channels in each hemisphere, when in competition with distracting 4-band NV speech (red) or 2-band NV speech (blue). Error bars indicated standard error of the mean.

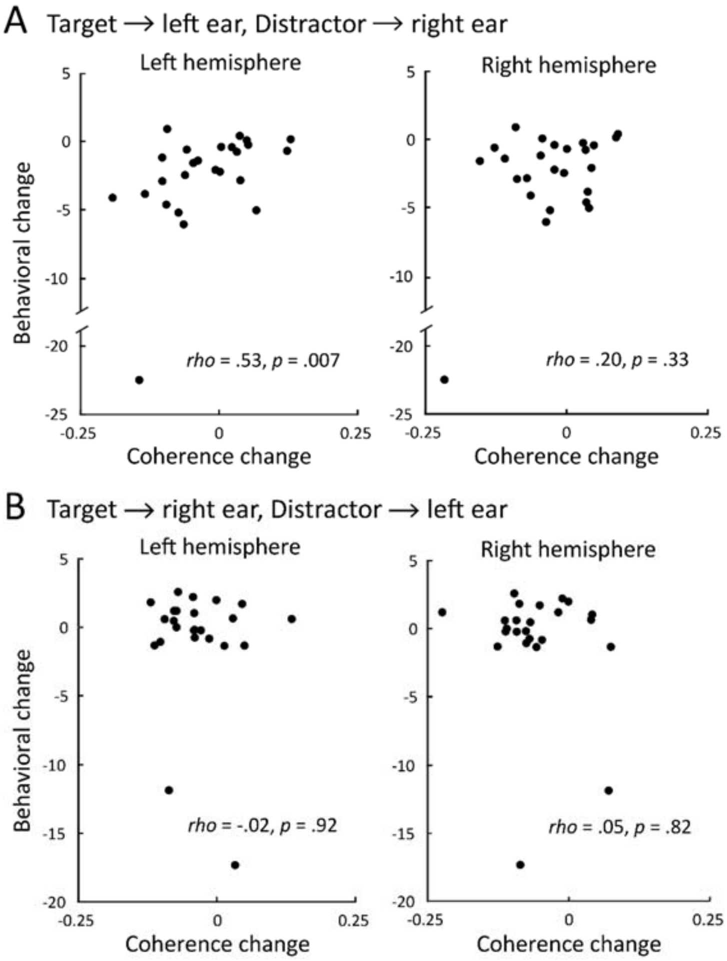

**Figure 5. Correlation between the strength of delta entrainment to target speech and the target speech intelligibility**. Inter-individual correlation between the change in target-brain coherence in the delta range before and after training (Coherence change) and the change in target speech intelligibility before and after training, when (A) target speech was delivered into left ear; (B) Target speech was delivered into right ear. Each dot corresponds to one participant. Note that the pattern of results (significant correlation in only the top left panel) did not change after excluding the outliers.

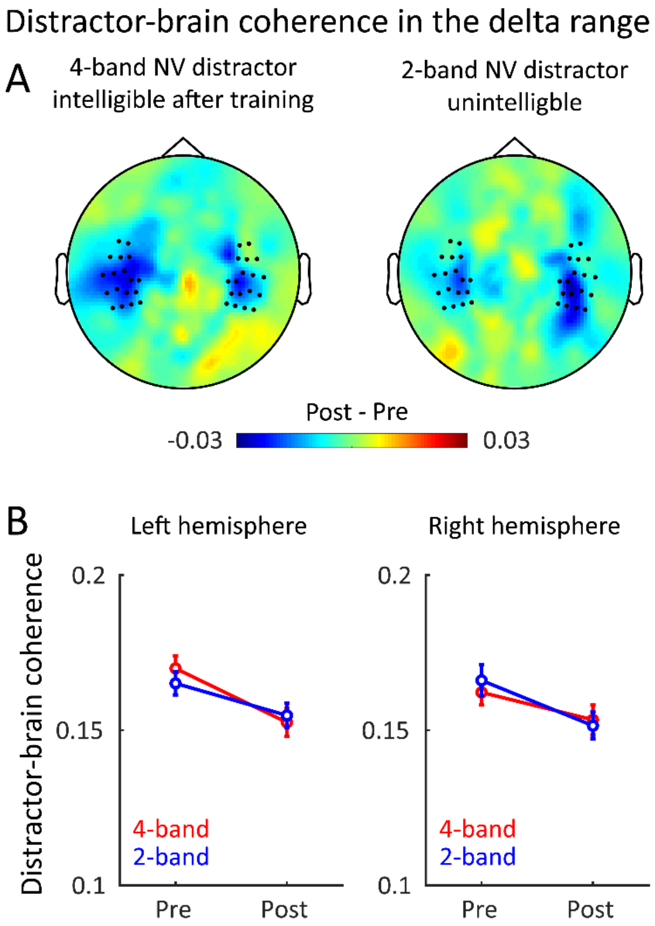

**Figure 6. Delta entrainment to distracting speech was not modulated by its intelligibility**. (A) Topographies of delta entrainment changes (Post-training minus Pre-training). Delta entrainment to distracting speech reduced after training, this irrespectively of the distracting signal. (B) Delta entrainment averaged across selected channels in each hemisphere, when in competition with distracting 4-band NV speech (red) or 2-band NV speech (blue). Error bars indicated standard error of the mean.

## Supplementary figures

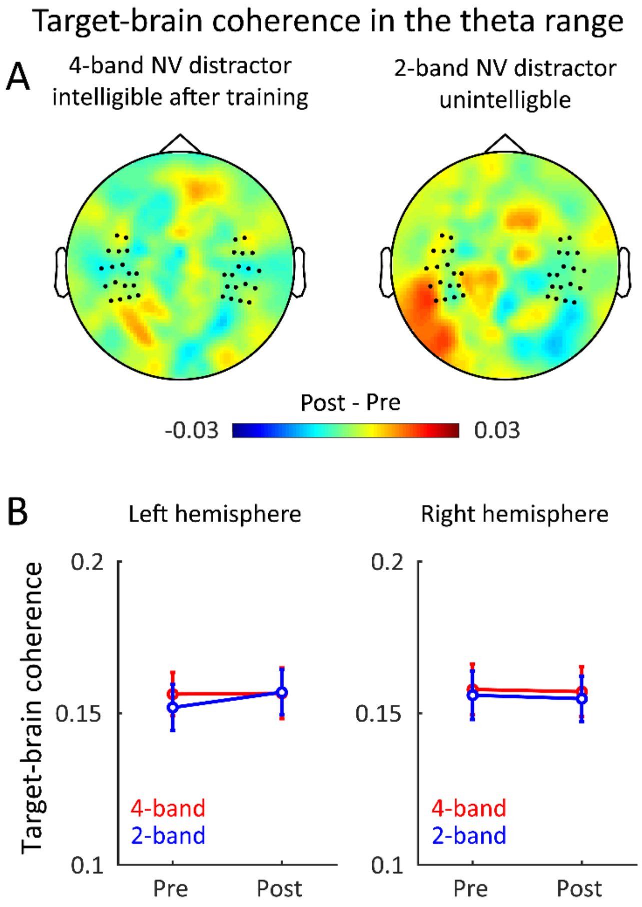

**Figure S1. Theta neural entrainment to target speech showed no change when the distracting NV speech gained intelligibility.** (A) Topographies of theta entrainment changes (Post-training minus Pre-training). No significant changes of training were observed for both 4- and 2-band NV conditions. (B) Theta entrainment averaged across selected channels in each hemisphere, when in competition with distracting 4-band NV speech (red) or 2-band NV speech (blue). Error bars indicated standard error of the mean.

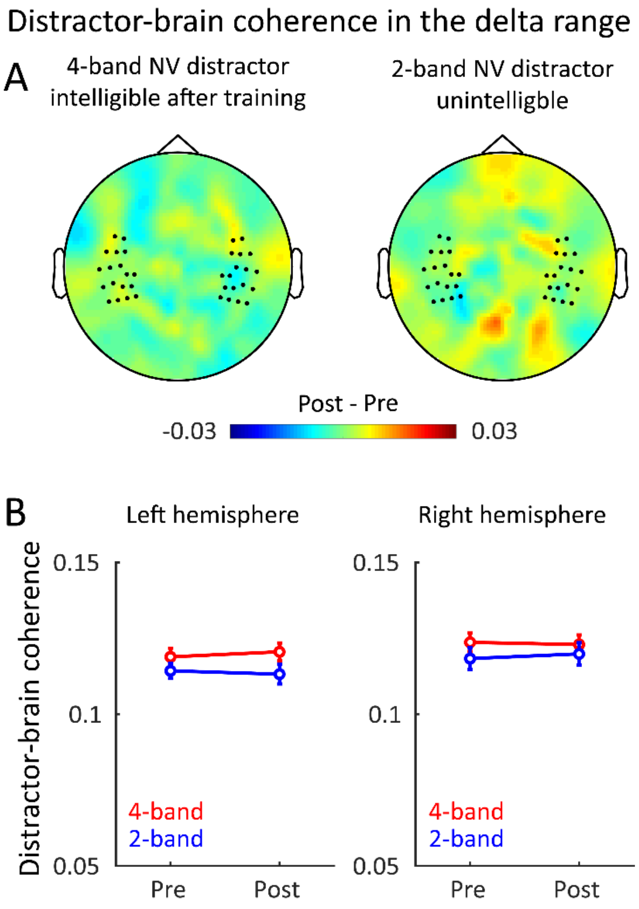

**Figure S2. Theta entrainment to distracting speech was not modulated by its intelligibility**. (A) Topographies of theta entrainment changes (Post-training minus Pre-training). No significant changes of training were observed for both 4- and 2-band NV conditions. (B) Theta entrainment averaged across selected channels in each hemisphere, when in competition with distracting 4-band NV speech (red) or 2-band NV speech (blue). Error bars indicated standard error of the mean.