HAL
open science

Behavioral Account of Attended Stream Enhances Neural Tracking

Moïra-Phoebé Huet, Christophe Micheyl, Etienne Parizet, Etienne Gaudrain

To cite this version:

Moïra-Phoebé Huet, Christophe Micheyl, Etienne Parizet, Etienne Gaudrain. Behavioral Account of Attended Stream Enhances Neural Tracking. Frontiers in Neuroscience, 2021, 15, pp.674112. 10.3389/fnins.2021.674112 . hal-03639824

HAL Id: hal-03639824

https://hal.science/hal-03639824v1

Submitted on 13 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 13 December 2021
doi: 10.3389/fnins.2021.674112

Behavioral Account of Attended Stream Enhances Neural Tracking

Moïra-Phoebé Huet[1,2†], Christophe Micheyl[3], Etienne Parizet[1] and Etienne Gaudrain[2,4]*

[1] Laboratoire Vibrations Acoustique, Institut National des Sciences Appliquées de Lyon, Université de Lyon, Villeurbanne, France, [2] CNRS UMR 5292, INSERM U1028, Auditory Cognition and Psychoacoustics Team, Lyon Neuroscience Research Center, Lyon, France, [3] Starkey France S.a.r.l., Creteil, France, [4] Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

OPEN ACCESS

Edited by:
Nima Mesgarani,
Columbia University, United States

Reviewed by:
István Winkler,
Research Centre for Natural Sciences, Hungarian Academy of Sciences (MTA), Hungary
Hamish Innes-Brown,
Eriksholm Research Centre, Denmark

***Correspondence:**
Moïra-Phoebé Huet
mphuet@jhu.edu

†Present address:
Moïra-Phoebé Huet,
Laboratory for Computational Audio Perception, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, United States

Specialty section:
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

Received: 28 February 2021
Accepted: 11 October 2021
Published: 13 December 2021

Citation:
Huet M-P, Micheyl C, Parizet E and Gaudrain E (2021) Behavioral Account of Attended Stream Enhances Neural Tracking. Front. Neurosci. 15:674112. doi: 10.3389/fnins.2021.674112

During the past decade, several studies have identified electroencephalographic (EEG) correlates of selective auditory attention to speech. In these studies, typically, listeners are instructed to focus on one of two concurrent speech streams (the "target"), while ignoring the other (the "masker"). EEG signals are recorded while participants are performing this task, and subsequently analyzed to recover the attended stream. An assumption often made in these studies is that the participant's attention can remain focused on the target throughout the test. To check this assumption, and assess when a participant's attention in a concurrent speech listening task was directed toward the target, the masker, or neither, we designed a behavioral listen-then-recall task (the Long-SWoRD test). After listening to two simultaneous short stories, participants had to identify keywords from the target story, randomly interspersed among words from the masker story and words from neither story, on a computer screen. To modulate task difficulty, and hence, the likelihood of attentional switches, masker stories were originally uttered by the same talker as the target stories. The masker voice parameters were then manipulated to parametrically control the similarity of the two streams, from clearly dissimilar to almost identical. While participants listened to the stories, EEG signals were measured and subsequently, analyzed using a temporal response function (TRF) model to reconstruct the speech stimuli. Responses in the behavioral recall task were used to infer, retrospectively, when attention was directed toward the target, the masker, or neither. During the model-training phase, the results of these behavioral-data-driven inferences were used as inputs to the model in addition to the EEG signals, to determine if this additional information would improve stimulus reconstruction accuracy, relative to performance of models trained under the assumption that the listener's attention was unwaveringly focused on the target. Results from 21 participants show that information regarding the actual – as opposed to, assumed – attentional focus can be used advantageously during model training, to enhance subsequent (test phase) accuracy of auditory stimulus-reconstruction based on EEG signals. This is the case, especially, in challenging listening situations, where the participants' attention is less likely to remain focused entirely on the target talker. In situations where the two competing voices are

clearly distinct and easily separated perceptually, the assumption that listeners are able to stay focused on the target is reasonable. The behavioral recall protocol introduced here provides experimenters with a means to behaviorally track fluctuations in auditory selective attention, including, in combined behavioral/neurophysiological studies.

Keywords: neural tracking, attentional switches, temporal response function (TRF), speech-on-speech, vocal cues

INTRODUCTION

Popularized by Cherry (1953) as the "cocktail-party problem" over 60 years ago, the question of how human listeners selectively attend a speaker amid one or several other concurrent voices, has attracted considerable interest to this day. While recent developments in machine-learning algorithms now allow machines to compete with – and in some situations, overtake – humans in this ability, a complete account of the psychological and neurophysiological processes at play remains elusive. Nonetheless, during the past decade, significant progress toward elucidating brain-activity correlates of the perceptual experience of listening selectively to one of two concurrent voices has been achieved. In particular, researchers have been able to identify features of electrically or magnetically recorded cortical signals which, after mathematical transformation, exhibit greater correlation with features of the target voice, than with features of the competing, non-target voice (e.g., Ding and Simon, 2012; Mesgarani and Chang, 2012; O'Sullivan et al., 2015).

One limitation of most earlier studies using the concurrent voice paradigm to study neural correlates of selective auditory attention, however, stems from their use of an experimental design in which participants were asked to attend to the target voice, and ignore the concurrent voice, over prolonged periods – from a few minutes to several tens of minutes. The premise that human listeners are able to unwaveringly maintain their auditory attention focused on a single sound source, be it a human voice, for such long time periods is at odds with introspective experience while participating in such – somewhat artificial – listening experiments involving concurrent voices. Our own experience, and informal reports from participants, strongly suggest that despite one's best efforts to stay focused on the target voice, the competing voice occasionally grabs one's attention. Unless such occasional attentional shifts can be controlled for, they can adversely impact data-analysis methods used to assess neural representations of the attended voice. Specifically, temporal response functions (TRFs) are obtained by relating temporal sequences of stimuli to the continuous brain activity recorded in response to them, by means of machine learning methods. Typically, for two competing speech streams, the temporal envelope of either of the streams is used to either predict the brain activity (*forward* TRF), or to be predicted by the brain activity (*backward* TRF). One of the differences between these two approaches is that forward TRFs treat each neural response channel independently while backward TRFs exploit the whole neural data in a multivariate context (Crosse et al., 2016). In addition, backward approaches can predict or "decode" which of the speakers the listener is attending. For this reason, backward TRFs are often called "*decoders*" whereas the accuracy to classify which speaker is attended is commonly referred as "*decoding accuracy*." Misestimating which of the competing streams is actually attended can impact the accuracy of these decoding algorithms in two ways. First, they can interfere with the training of the algorithm, if the brain responses used for training span epochs during which the participant was actually attending to the non-target voice. Second, they can interfere with the measured decoding accuracy of the algorithm at test-time, if the brain responses on which the algorithm is tested are assumed to contain only target-attend, or only non-target-attend, epochs.

To mitigate this issue, some investigators have made attempts to assess the occurrence of attentional shifts during the experiment. For instance, O'Sullivan et al. (2015) asked their listeners multiple-choice questions following every 1-min stimulus, to check that the listener had been paying attention to the target story. One limitation of this approach, however, is that listeners may have been able to answer the questions correctly, even if they did not always pay close attention to the target stream. Crosse et al. (2015) asked participants to press a button whenever they were listening to the target voice. However, it is possible that most listeners are unable to, simultaneously, perform the demanding listening task, and to accurately report their auditory-attention status accurately. Moreover, asking listeners to press buttons according to their attention while they are listening introduces a secondary task, which may perturb performance in the primary, selective-attention task. The problem of attention shifts has been acknowledged, and attempts to develop attention-decoding algorithms that can cope with such shifts have been developed (Akram et al., 2016, 2017; Miran et al., 2018; Jaeger et al., 2020). However, except when a distraction was purposely inserted (Holtze et al., 2021), most studies in the literature seem to have remained limited by the almost complete lack of detailed data regarding the timing of attentional shifts in selective-listening experiments with concurrent voices.

The issue of attentional switching across two concurrent auditory streams can hardly be approached without considering how easy, or hard, it is for listeners to perceptually separate these two streams. Previous studies have shown that two of the most important cues used by human listeners to separate concurrent voices are spatial separation and differences in the fundamental-frequency (F0) or timbre of these voices (Bronkhorst, 2015; Middlebrooks, 2017). Recently, two studies have investigated attentional switching with spatial cues. Bednar and Lalor (2020) showed that it was possible to reconstruct, with TRFs, the trajectory of attended and unattended moving sound sources. In the second paper, carried by Teoh and Lalor (2019), participants had to focus on a target voice while both talkers (target and

masker) were instantaneously alternated between the left and right ears. The authors showed that it was possible to significantly improve the auditory attention decoding accuracy with the inclusion of spatial information.

In the present work, we investigate the assumption that listeners are able to maintain their attention focused on the target speech stream and assess when a participant's attention was directed toward the target, the masker, or neither with a test that was designed to provide experimenters with a means of inferring fluctuations in auditory selective attention: the Long-SWoRD test (Huet et al., 2021). Here, participants' answers are used to infer, retrospectively, when they were listening to the target, or to the masker. The difficulty of the task, and hence the likelihood of an increase in attentional switch to occur during the course of the stories, was modulated with vocal cues. The attention course was modeled by combining the participants' responses with three different parameters. These parameters, described in section "'Inferred Stimuli," depict the speed and duration of attentional switches as well as the actual sound source that receives the focus of attention. This better representation of actual attended speech is expected to yield a better stimulus-reconstruction evaluation since it takes into account attentional dynamics.

MATERIALS AND METHODS

Participants

Twenty-one participants, aged between 19 and 25 (μ = 21 years, σ = 1.76), participated in the experiment. All of them were native French speakers and had audiometric thresholds $\leq$ 30 dB HL at audiometric test frequencies between 125 Hz and 8 kHz. Participants gave informed consent before taking part in the study and were paid an hourly wage for participation.

Procedure

The Long-SWoRD test (Huet et al., 2021) was used to obtain estimates of the attended stream at different time points of the stimulus. Two competing stories were presented diotically at the same time to both ears. Participants were instructed to focus on one of the two concurrent speech streams (the "target"), while ignoring the other (the "masker"). At the end of the trial, nine keywords, arranged in a three-by-three matrix, were presented on a screen facing the participant. The three rows corresponded, from top to bottom, to the beginning, middle and end portions of the story. Each row included, in a random order, one keyword from the target sentence, one keyword from the "masker" sentence, and an "extraneous" keyword which was contained neither in the target nor in the masker sentence. Participants were instructed to select the three keywords in the target story with the constraint that they could select only one keyword in each row.

The difficulty of the task, and therefore the probability of attentional switches occurring, was modulated by manipulating the perceptual distance between the two competing stories in terms of voice (see section "Voice Manipulation"). The experiment was arranged into 12 blocks, randomly distributed between three levels of difficulty. Within each block, there were 12 trials and the same distance between the target and the masker voices was kept. The characteristics of these voices are described in the next section.

Data collection lasted 60–100 min, and the entire procedure was completed in a single session. Participants were instructed to avoid eye movements to reduce potential noise in the electroencephalographic (EEG) recording. Stimuli were presented with OpenSesame (Mathôt et al., 2012). Participants listened to stimuli diotically over Sennheiser HD250 Linear II headphones in a sound-attenuated booth. The presentation level was calibrated to 65 dB SPL using an AEC101 artificial ear and sound level meter LD824 (Larson Davis, Depew, NY, United States).

Stimuli

Material Content

This material was previously developed and used in two behavioral studies (Huet et al., 2018, 2021). Short, interesting and engaging stories, extracted from the French audiobook "*Le Charme discret de l'intestin*" (The Inside Story of Our Body's Most Underrated Organ) (Enders et al., 2015), provided the stimulus set for the target and masker streams. Each story was composed of 1–5 sentences. Each trial, composed of a target story and a masker story of similar length, lasted between 11 and 18 s.

The three target keywords that participants would later have to identify were selected at three different times within the story: one keyword near the beginning of the story, one keyword toward the middle of the story, and one keyword toward the end. The same selection procedure was applied for the masker keywords, whereas the extraneous keywords originate from other trials. Further details and considerations about the choice of keywords and statistical analyses of the linguistic features of the stimuli can be found in Huet (2020) and Huet et al. (2021).

Voice Manipulation

Manipulating the parametric distance in semitones (st) between the target and the masker voices and thus, varying the difficulty level of the task is an approach used in previous experiments (e.g., Darwin et al., 2003; Vestergaard et al., 2009; Ives et al., 2010; Başkent and Gaudrain, 2016). The audio stimuli were originally recorded by an adult female speaker. This original voice, analyzed and resynthesized without modification (i.e., with a voice distance of 0 st) with the STRAIGHT toolbox (Kawahara et al., 1999) implemented in MATLAB, was chosen as the target voice. For the creation of the three masker voices, the voice pitch (F0) and vocal-tract length (VTL) were then manipulated during the analysis-resynthesis. The total distance between the target and the masker voices is then calculated, in semitones, as $\sqrt{\triangle F0^2 + \triangle VTL^2}$. Total distance values of 1.14 st, 3.42 st, and 5.13 st were chosen to constitute three levels of masking, difficult, intermediate, and easy, respectively. In a previous experiment (Huet et al., 2021), we were able to estimate that a difference of 5.13 st is a good control condition since the participants made almost no error as a result of the large voice difference. In addition, participants did not make more masker errors than extraneous errors, suggesting no target-masker confusions and

TABLE 1 | Distance between the target and the masker voices, in semitones.

Condition	ΔF0	ΔVTL	$\sqrt{\Delta F0^2 + \Delta VTL^2}$
Difficult	−1.6	0.4	1.14
Intermediate	−3.2	1.21	3.42
Easy	−4.8	1.82	5.13

no or almost no attentional switches. Parameter values for the three masker voices are provided in **Table 1**.

"Inferred" Stimuli

"Inferred stimuli" are reconstructed stimuli derived from the original stimuli and from the behavioral responses aiming at estimating the actual attended stream, switching between target and masker streams. Participants' responses provide information at three key moments in the story (beginning, middle, and ending keywords). It is important to note that there are not just three keywords that provide information, but six: three target and three masker keywords. For each response, the participants cannot choose both target and masker, but are faced with a choice. Thus, if a participant selects the masker keyword in a line of the matrix, it is possible to hypothesize that the subject was not listening to the associated target keyword. The information is therefore not limited in time to the chosen keyword, but also extended to the associated non-selected keyword in the other stream, and which may not be occurring exactly at the same time. Thus, there are three key moments in the stories, bound by the time limits of the target and the masker keywords, which provide information. For convenience, these key moments will be named "windows." For instance, in **Figure 1A**, the first target and masker keywords overlap while conversely, the second target and masker keywords are separated by 1 s. Therefore, these key moments, or windows, can have varying durations. In addition, the windows duration started at the beginning of the keyword that appears first, and ended at the end of the keyword that ended last. Based on the participants' answers, the inferred stimuli were modeled following various strategies differing in how three aspects of the task were handled: the duration and speed of attentional switches (described respectively in sections "Extrapolation of Attentional Scope" and "Attentional Switch Speed") as well as the sound source to which attention is focused (described in section "Extraneous Keywords Fillers").

Mathematically, the inferred stimulus can be expressed as follows:

$$\hat{x}(t) = \begin{cases} f(x_T, x_M, R_i)(t) \text{ if } t \in \text{ keyword window } i \\ g(x_T, x_M)(t) \text{ otherwise} \end{cases}$$

Where x_T and x_M are the target and masker stimuli, respectively; R_i designates the response to keyword i; and f is the following function:

$$f(x_T, x_M, R_i) = \begin{cases} x_T \text{ if } R_i \text{ is target keyword} \\ x_M \text{ if } R_i \text{ is masker keyword} \\ h(x_T, x_M) \text{ otherwise} \end{cases}$$

The functions g and h are defined below.

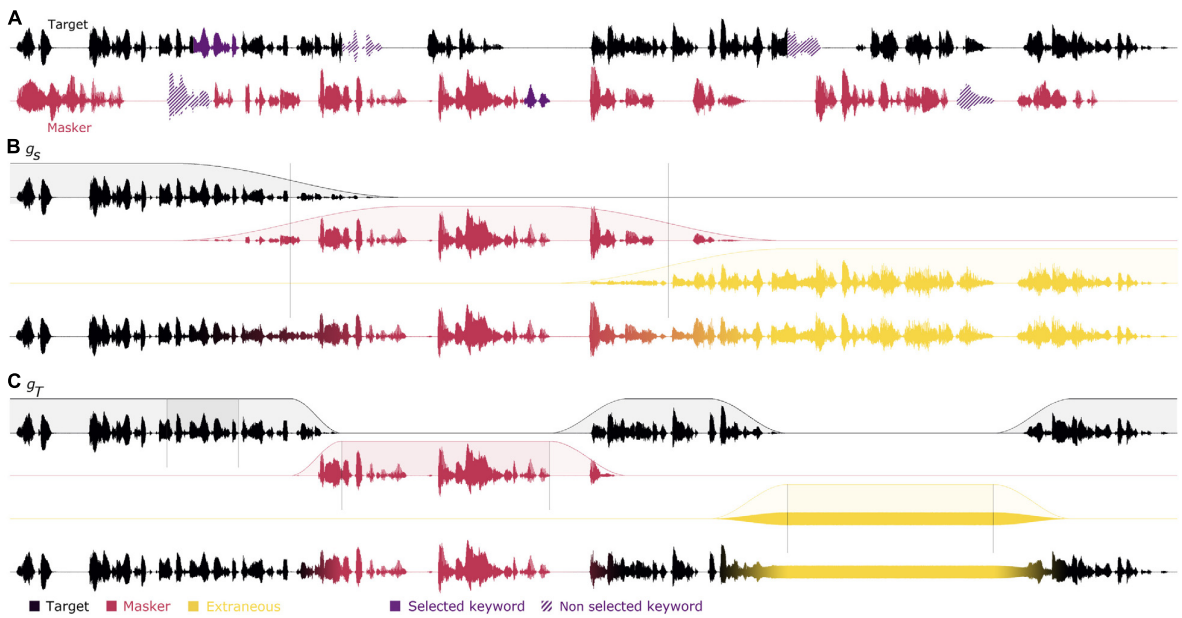

FIGURE 1 | Creation steps of inferred stimuli. Panel **(A)** represents an example of a trial where the target is in black and the masker is in red. In this example, the participant has answered the first target keyword (highlighted in purple), the second masker keyword and the third extraneous keyword. The unselected keywords are shown with a hatched pattern. **(B)** The three segments are built according to the participant's answer with an attention switch of 3 s and the extraneous segments filled with the mixture of the target and masker (i.e., method h_+, in yellow). The attention scopes are interpolated with the segment method g_S. The three segments are then added together. **(C)** The three attention scopes are built according to the participant's answer with an attention switch of 1 s and the extraneous sections filled with noise (h_N). When no behavioral information is known, the attention scope is filled with the target (i.e., g_T). The three parts are then added together.

Extrapolation of Attentional Scope

Outside of the windows defined by the keyword positions, since there is no behavioral data collected, the attentional locus is not explicitly known and needs to be inferred. This situation occurs for instance at the very beginning of the stories (before the first keyword), at the end of the stories (after the last keyword), and between consecutive windows. Two different approaches were used to estimate attention outside of the windows, thereafter referred to as *attentional scopes*.

In the first approach "*segments*," illustrated in **Figure 1B**, the stimulus duration was divided into three segments based on the temporal positions of the keywords. The cut-out points between segments were placed in time halfway between two consecutive windows, or at the beginning and end of the stimulus. In this case, we are making the assumption that the attention information provided in the windows remains over a wider duration than the window itself, dividing the unknown segments equally across windows. This corresponds to a function g as follows:

$$g_s\left(x_T, x_M\right)(t) \;=\; f\left(x_T, x_M, R_i\right)(t)\,,$$

where R_i is the closest known response.

In the second approach "*target windows,*" illustrated in **Figure 1C**, it is assumed that the information contained in the window should be limited to the window itself only, which is the opposite of the first approach, which extended the information to an entire segment. Between windows, it then can be assumed that the participant was always listening to the same stream (either the target or the masker). **Figure 1C** illustrates a situation where the participant listens to the target outside of the windows. The cases where the listener always listens to the target in-between keywords correspond to the following variant of the g function:

$$g_T\left(x_T, x_M, R_i\right) \;=\; x_T$$

This parameter will be referred to as "scope" with the function g_S "segments" and the function g_T "target windows."

Extraneous Keywords Fillers

In addition to attending to the target, or to the competing (non-target) voice, participants in concurrent-voice experiments may also, at times, not be attending to either. Thus, in addition to inferring when attention was directed to the target or to the masker, it is also important to try to infer when it is not. To this aim, participants' selection of displayed keywords that belonged to neither of the two stories played during the trial, i.e., the extraneous keywords, is instrumental. The selection of an extraneous keyword over a keyword from either story may be an indication that, when the keywords that the participant failed to select were presented, the participant was not attending to either of the two stories being played. In such a case, neither the target nor the masker is more appropriate than the other stream to represent the attended stimulus. One way of representing this situation thus consists in using the mixture of the two streams as attended stimulus (illustrated in **Figure 1B**). This, corresponds to a function h as follows:

$$h_{+(x_T,x_M)} = x_T + x_M$$

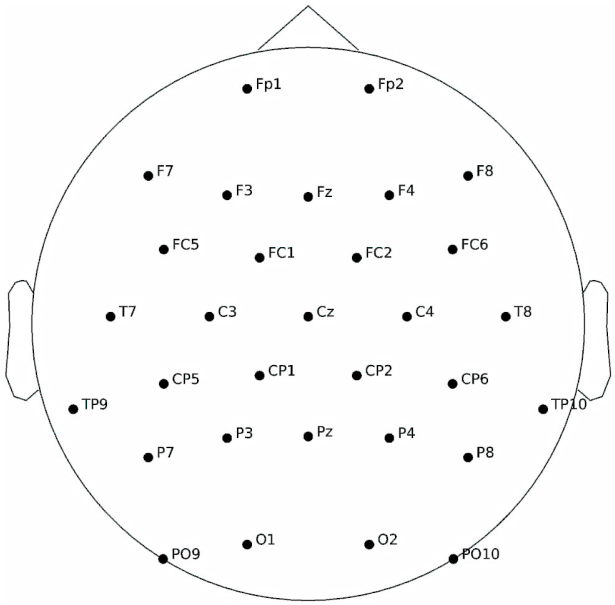

FIGURE 2 | Electroencephalographic channel positions.

This might account for situations where the listeners were actually dividing their attention between the two streams, which led them to fail to recall the corresponding keyword at the end of the trial. However, it is also possible that the listeners, in these situations, were actually not attending *any* of the presented streams. In these situations, the mixture does not seem the most appropriate acoustic correlate of what the participant is focusing on, and instead, a random noise signal (illustrated in **Figure 1C**, noted h_N) has been used to fill in the extraneous keywords. Additionally, a target speech story from another trial of the Long-SWoRD corpus (noted h_S), randomly selected, was also used as a control speech signal. The root-mean-square level of the extraneous fillers have been adapted to match the root-mean-square level of the target. Finally, for completeness, we also considered the case where extraneous keyword responses would be treated as the target (h_T) or masker (h_M) streams:

$$\begin{aligned} h_T\left(x_T, x_M\right) &= x_T \\ h_M\left(x_T, x_M\right) &= x_M \end{aligned}$$

This parameter will be referred to as "filler."

Attentional Switch Speed

The third parameter is the speed with which participants can switch from one voice to another. The duration of this attentional switch is modeled as the slope of the edges of the time windows. Three values were used, 1, 2, and 3 s, implemented as raised-cosine ramps. Those values were chosen as they could capture attentional switches: slower than speech modulations, but shorter than sentences to limit overlap across segments. This parameter will be referred to as "speed."

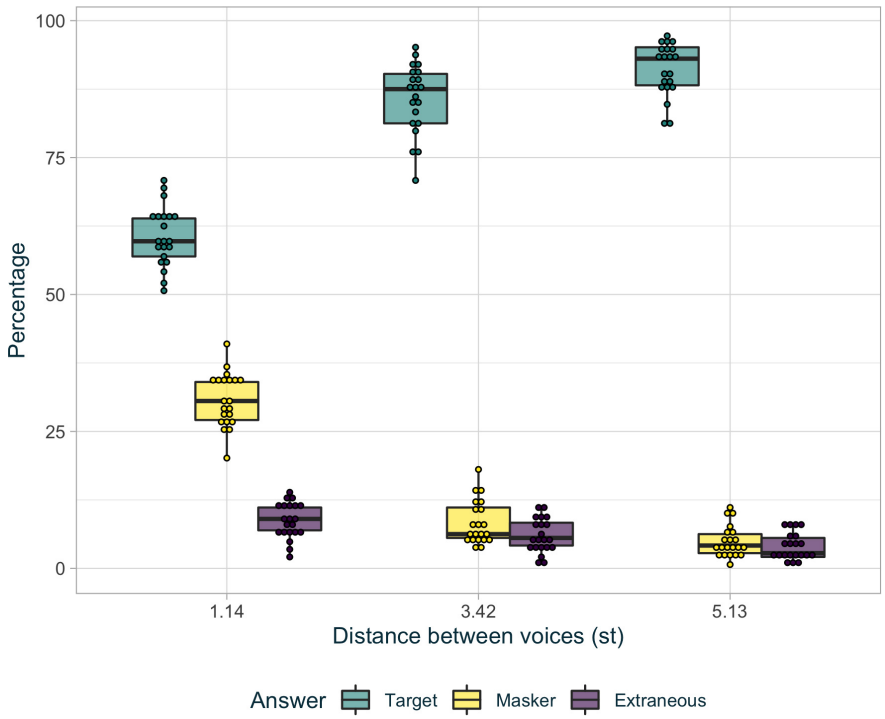

FIGURE 3 | Percentage for target answers (in green), masker answers (in light yellow), and extraneous answers (in dark purple) in each level of difficulty. The points represent each identified keywords percentage for every participant in each condition. The hinges of the boxplot represent the first and the third quartile. The middle of the boxplot is the median. The whiskers extend up to 1.5 times the interquartile range.

Data Acquisition and Signal Processing

Electroencephalographic data were recorded using an ActiCap (Brain Products, Munich, Germany) with a setup of 31 channels at a sampling rate of 1000 Hz (see **Figure 2** for more information). A trigger was sent at each start of a new trial on a parallel port with a precision of 1 ms. EEG data were then band-pass filtered between 2 and 8 Hz. Finally, to decrease processing time, EEG data were downsampled to 64 Hz.

The stimulus speech envelope was extracted with a gammatone filterbank (Søndergaard et al., 2012; Søndergaard and Majdak, 2013) followed by a power law according to Biesmans et al. (2017). The gammatone filterbank was composed of 28 bands centered on frequencies from 50 to 5000 Hz, equally spaced on the ERB_N scale (Glasberg and Moore, 1990). The envelopes of each frequency band were extracted by taking the absolute value and then raising it to the power of 0.6. A single envelope for the stimulus was then computed by averaging the 28 envelopes. The speech envelope was then downsampled to 64 Hz and low-pass filtered below 8 Hz, following the method described by O'Sullivan et al. (2015).

Backward Modeling and Stimulus-Reconstruction

Regularized linear regression was employed to relate the neural data to the envelopes and the decoders were calculated using the MNE-Python library (Gramfort et al., 2014), using the backward method, i.e., reconstructing the audio envelope from the EEG recording. These decoders, equivalent to backward TRFs, are composed of weights that can be estimated by a linear regression for a set of N electrodes at different lags t. In this experiment, we investigated time lags between −900 to 0 ms (meaning the audio could precede the EEG up to 900 ms) by steps of 1 sample of the EEG recording (15.6 ms). Therefore, the EEG data were cut according to the duration of the trials added with the lags t. As for speech envelopes, they were padded with zeros to match the number of samples of EEG data. Finally, the reconstruction of the speech envelope $\hat{S}_t$can be obtained as follows:

$$\hat{S}_t = \sum_{n=1}^{N} \sum_{\tau} d_{\tau,n} R_{t-\tau,n}$$

Where **R** represents the matrix that contains the shifted neural responses of each electrode n at time $t = 1 \ldots T$. A ridge regression was used to obtain the weights of the *decoder* d as follows:

$$d = \left(RR^{\mathrm{T}} + \lambda \mathrm{I}\right)^{-1} \left(RS^{\mathrm{T}}\right)$$

where λ is the regularization parameter, chosen to optimize the stimulus-response reconstruction, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{S}$ is the envelope of the speech signal. The optimal ridge parameter was fit with an adaptive procedure according to Crosse et al. (2016) and set to $10^{1/2}$. Decoders were estimated per trial, for each subject in each condition. The stimulus-reconstruction of a single trial was predicted in a leave-one-out

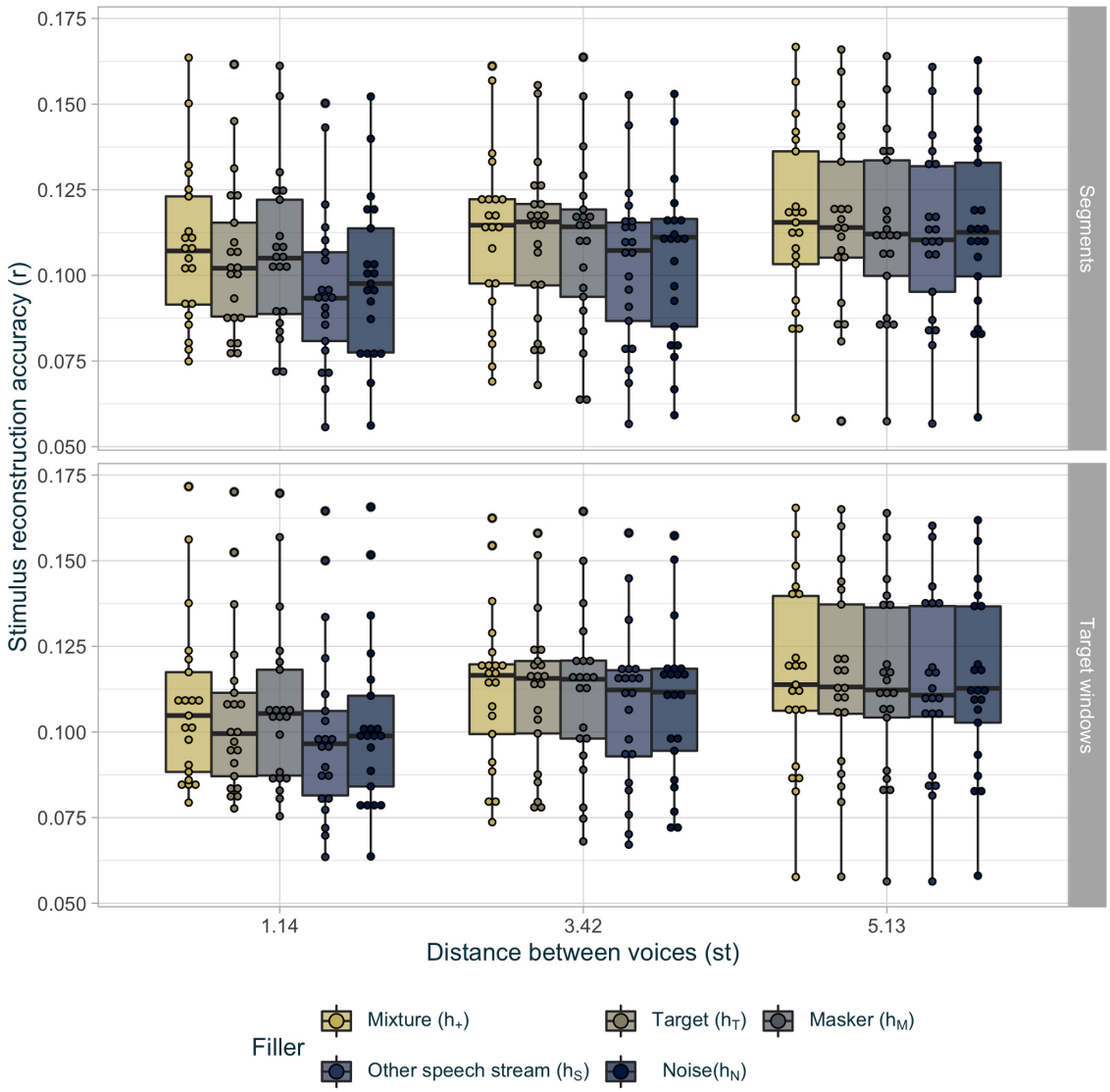

FIGURE 4 | Average reconstruction accuracy *r* for inferred envelopes per condition. The points represent the scores for every participant in each voice condition for the extraneous keyword (in color) and the attentional scope (top and bottom). The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The whiskers extend over 1.5 times the interquartile range.

fashion. To be more precise, each subject had 48 trials per condition. Each trial was reconstructed with the averaged decoder trained on the 47 other trials. The stimulus-reconstruction was evaluated with the Pearson's correlation coefficient between the reconstructed speech envelope and the original speech envelope. This reconstruction accuracy is thereafter noted *r*. The temporal resolution of the reconstructed envelope was the same as that of the original envelope (64 Hz).

Statistical Analyses

All statistics were performed using R (R Core Team, 2017). All the linear mixed models (LMMs) were implemented with the *lme4* package (Bates et al., 2014). The models were implemented using a top-down strategy on data (Zuur et al., 2009). The final model is reported with the *lme4* syntax such as Equation 1:

$$\text{Score} \sim \text{factor}_\text{A} \times \text{factor}_\text{B} + \left(\text{factor}_\text{A} \times \text{factor}_\text{B} | \text{subject}\right) \quad (1)$$

The full-factorial model is indicated by the fixed effect term $factor_A \times factor_B$ and includes main effects and interactions for these two main conditions. The last term of the equation describes an individual random intercept and slope per subject for $factor_A$ and $factor_B$. $Factor_A$, $factor_B$, and *Score* will be specified in section "Results" for each analysis.

For an easier interpretation, the *afex* package (Singmann et al., 2019) was used to compute the statistics of main effects. To do so, the final model was compared to restricted models in which the effect estimated is fixed and equal to zero. Finally,

post hoc analyses were computed with a false discovery rate (FDR) correction (Benjamini and Hochberg, 1995).

RESULTS

Behavioral

Figure 3 shows the percentage of identified keywords ("target," "masker," and "extraneous") for each level of voice difficulty (difficult: 1.14 st; intermediate: 3.42 st; and easy: 5.13 st). A generalized linear mixed model (gLMM) was fitted on the binary (correct/incorrect) scores. Such models are well suited to preserve homoscedasticity and to minimize the effects of saturation in binomial data. For each keyword within each trial, if the participant selected the target keyword, the score was positive (i.e., score correct = 1). On the other hand, if the participant selected the masker keyword or the extraneous keyword, the score was considered incorrect (i.e., score incorrect = 0).

Equation 2 shows the final model with a top-down strategy modeling:

$$\text{Score} \sim \text{voice} + (\text{voice}|\text{subject}) \quad (2)$$

Participants had better scores when the distance between the target and the masker voices was larger ($2.13, SE = 0.12, z = 18.02, p < 0.001$). *Post hoc* analysis showed that average scores in each voice condition were all different from one another. In addition, there were significantly more masker responses than extraneous responses when stimuli were only presented with the 1.71 st voice ($z = 13.06, p < 0.001$), but not for a voice distance of 3.42 st ($z = 1.24, p = 0.22$), or 5.13 st ($z = 0.09, p = 0.93$). These results indicate that participants were listening, at least partially, to the masker voice instead of the target voice only in the difficult condition whereas the switches between target and masker were limited in the two other conditions.

Stimulus-Reconstruction Evaluation

Modeling Parameters

Because the parameter space defining the possible inferred stimuli is rather larger, before comparing it to the original approach, we selected the set of parameters that gave the best reconstruction.

Figure 4 shows stimulus reconstruction accuracy in each voice condition as a function of extrapolation method (scope: segments g_S or target windows g_T) and treatment method for the extraneous keywords (filler : mixture h_+, target h_T, masker h_M, and other speech stream h_S or noise h_N), averaged across speed of attentional switch (speed: 1, 2, or 3 s). The influence of the three modeling parameters (scope, filler, and speed) as well as the voice distance (difficult: 1.14 st; intermediate: 3.42 st; and easy: 5.13 st) was also analyzed with a LMM fitted to the Fisher transformed Pearson's correlation r values representing the reconstruction accuracy. Equation 3 indicates the final model:

$$\text{r} \sim \text{voice} \times \text{filler} + \text{scope} + (1|\text{ subject}) \quad (3)$$

Similarly to the analysis in the previous section, the distance between voices had an effect on reconstruction performance ($\chi^2(2) = 282.16, p < 0.001$) but post-doc analyses with a FDR correction did not identify any individual difference between the conditions (see **Table 2**). Regarding the modeling parameters, the filler for the extraneous keyword (h_T, h_M, h_+, h_N, or h_S) had an effect on the reconstruction of the inferred stimuli ($\chi^2(4) = 100.59, p < 0.001$). *Post hoc* analyses, detailed in **Table 2**, showed that when the time segment corresponding to extraneous keyword responses was filled-up with the mixture (h_+), stimulus reconstruction accuracy was the best. The target (h_T) and masker (h_M) streams were second-best, followed by the noise (h_N) and, lastly, the other speech (h_S) stream. The interaction between the latter two factors ($\chi^2(8) = 25.17, p < 0.01$) showed an effect of the filler, only when the distance between the two voices was 1.14 st ($\chi^2(3) = 78.04, p < 0.001$) and 3.42 st ($\chi^2(3) = 38.38, p < 0.001$) but not for 5.13 st ($\chi^2(3) = 8.71, p = 0.07$). Furthermore, in this easiest voice condition (5.13 st), the reconstruction accuracy performed with the target (h_T) as the filler reached the reconstruction accuracy performed with the mixture (h_+) $\left[t(20) = 0.29, p = 0.77\right]$ while the reconstruction accuracy performed with the masker (h_M) was equivalent to the reconstruction accuracy performed with the noise (h_N) ($t(20) = -0.39, p = 0.74$). Finally, the attentional scope (g_T and g_s) also had a significant effect ($\chi^2(1) = 9.23, p < 0.01$), with a better performance when the target-windows approach (g_T) was used over the three segments (g_s) ($t(20) = -3.13, p < 0.01$). It is noteworthy than the attentional switch (1, 2, or 3 s) speed had no effect on reconstruction performance [$\chi^2(2) = 0.45, p = 0.50$].

In conclusion, the best stimulus-reconstruction evaluation was obtained when the behavioral response was used only at the location of the keywords and the remaining segments were filled with the target stream (g_T), while the mixture was used in case of extraneous keyword responses (h_+), regardless of the attentional switch duration. In the following section, the term "behavioral decoder" denotes a decoder obtained using these best parameters (g_T, h_+) and a 2-s attentional switch duration.

Target vs. Inferred Stimulus

In this section, the best behavioral decoder is compared with the original target decoder. The evaluation of these two decoders was

TABLE 2 | *Post hoc* analyses for Equation 3.

Main effect	Individual comparison	Statistics
Difference between voices	1.14 st vs. 3.42 st	$t(20) = -1.82, p = 0.13$
	1.14 st vs. 5.13 st	$t(20) = -2.52, p = 0.06$
	3.42 st vs. 5.13 st	$t(20) = -1.17, p = 0.26$
Extraneous keyword filler	Mixture vs. target	$t(20) = 5.26, p < 0.001$
	Mixture vs. masker	$t(20) = 5.43, p < 0.001$
	Target vs. masker	$t(20) = 0.65, p = 0.53$
	Target vs. other speech stream	$t(20) = 8.11, p < 0.001$
	Target vs. noise	$t(20) = 6.54, p < 0.001$
	Other speech stream vs. noise	$t(20) = -6.64, p < 0.001$

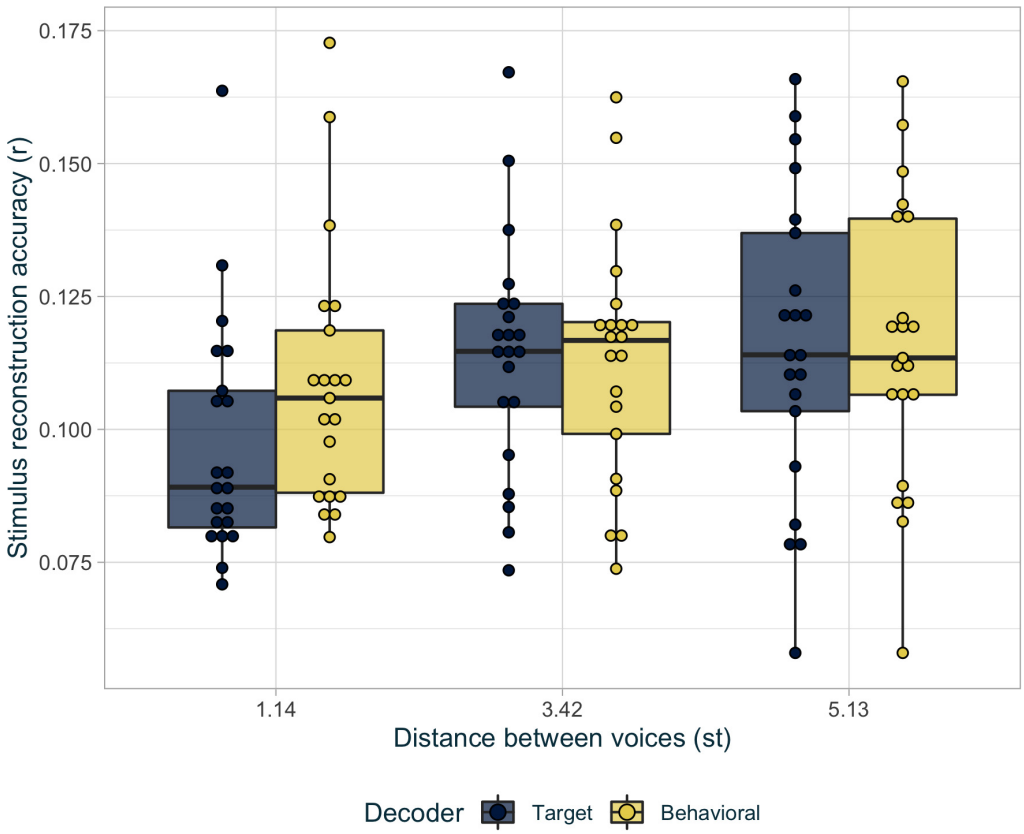

FIGURE 5 | Average reconstruction accuracy *r* per condition for the decoder based on original target stimuli (in dark blue), and for the decoder based on the best inferred stimuli (in light yellow).

analyzed with a LMM fitted on the Fisher-transformed *r* values representing reconstruction accuracy:

$$\mathrm{r} \sim \mathrm{voice} \times \mathrm{decoder} + (\mathrm{voice}|\ \mathrm{subject}) \tag{4}$$

Based on likelihood-ratio tests, stimulus reconstruction accuracy depended significantly on the distance between the two voices ($\chi^2(2) = 11.95, p < 0.01$), on whether the target or behavioral decoder was used ($\chi^2(2) = 11.1, p < 0.001$), and the interaction between these two factors ($\chi^2(2) = 28.85, p < 0.001$). In *post hoc* comparisons, the behavioral decoder was significantly superior to the original target decoder in the most challenging voice condition, while there was no difference between decoders for the two easier voice conditions (see **Figure 5** and **Table 3**). In addition, unlike reconstruction with the target decoder, there was no difference in performance between the voice conditions when reconstructing with inferred stimuli (see **Table 3**).

TABLE 3 | *Post hoc* analyses for the Equation 4 interaction.

Individual comparison	Statistics
1.14 st: target vs. behavioral decoder	$t(20) = -4.6, p < 0.001$
3.42 st: target vs. behavioral decoder	$t(20) = 0.76, p = 0.54$
5.13 st: target vs. behavioral decoder	$t(20) = 0.62, p = 0.54$
Target decoder: 1.14 st vs. 3.42 st	$t(20) = -4.55, p < 0.01$
Target decoder: 1.14 st vs. 5.13 st	$t(20) = -4.04, p < 0.01$
Target decoder: 3.42 st vs. 5.13 st	$t(20) = -0.54, p = 0.6$
Behavioral decoder: 1.14 st vs. 3.42 st	$t(20) = -1.4, p = 0.27$
Behavioral decoder: 1.14 st vs. 5.13 st	$t(20) = -1.64, p = 0.23$
Behavioral decoder: 3.42 st vs. 5.13 st	$t(20) = -0.64, p = 0.6$

DISCUSSION AND CONCLUSION

In this work, we assessed whether the reconstruction accuracy of attention-decoding algorithms in a selective-listening task can be enhanced by making use of information regarding the time-course of attentional shifts inferred using participants' answers in a keyword-recall task performed immediately after a selective-listening task. The answer to this question was found to be positive. Consistent with our hypotheses, an advantage of the new decoding method, including estimates of the timing of attention shifts, was only observed in challenging listening conditions, where the participants' attention was less likely to remain focused on the target talker throughout the entire listening-trial duration; no improvement over the simpler decoding algorithm, which did not make use of information regarding the timing of attentional shifts during the trial, was found for the easy listening conditions.

These findings are particularly relevant for future applications of the attention-decoding paradigm. While most experimental work thus far has focused on normal-hearing (NH) listeners attending to clearly separated speech streams (most often, two speakers of different genders presented dichotically), one of the ultimate goals of this line of research is to use attention decoding to enhance speech perception for hearing impaired (HI) listeners in challenging situations. However, HI listeners do not benefit as much as NH listeners from voice differences in competing speech (e.g., Festen and Plomp, 1990), and the situation seems even more severe for cochlear implant (CI) users (e.g., Pyschny et al., 2011; El Boghdady et al., 2019). In the present study, we show that, not only can stimulus reconstruction accuracy still be performed under conditions where voice cues are not salient, but it can also be improved further, to the point that decoding performance in difficult listening conditions can equal performance in easy listening conditions – provided that the decoder is trained with stimuli that account for the behavioral responses of the participant that indicate attention switches.

Lack of Benefit in Less Challenging Conditions

The lack of stimulus-reconstruction enhancements in easy conditions can be explained by a reduced number of errors from participants. Two to three times more errors were made in the challenging condition than in the other conditions. Since inferred stimuli are based on the participants' answers, in a trial where no error occurs in the behavioral task, the inferred stimulus is identical to the original target. It is therefore not surprising that under easy conditions, minimal differences between the inferred stimuli and the original target reconstructions were observed. However, it remains unclear whether this lack of errors truly reflects a lack of attentional switch between competing speech streams, or lack of sensitivity of the behavioral procedure used here; specifically, the procedure may have failed to capture momentary shifts in attention in-between keywords. Indeed, in connected speech, perceptual, and cognitive compensation, a process sometimes referred to as *phonemic restoration*, can help a listener infer missing segments (Bashford et al., 1992). It is possible that the participants' attention sometimes wavered away from the target, but that they still managed to infer the correct response in the task nonetheless. However, the Long-SWoRD test was designed such as to limit the possibility of such restoration mechanisms. First, the target and masker sentences both came from the same audiobook, and had largely overlapping lexical fields. In addition, the extraneous keywords were chosen to be equally likely to occur in the context of the target and the masker – see Huet et al. (2021), for a detailed analysis of the material. Given these methodological-design precautions, it seems less likely that phonemic restoration played a major role in compensating for momentary attention switches; it seems more likely that attention switches remained very limited.

Optimal Parameters for Inferred Stimuli

Several parameters were used to model the inferred stimuli. Extraneous keyword filling seems to be the most important factor, with an improved reconstruction in a challenging condition when the extraneous keyword is replaced by the mixture (h_+), the target (h_T) $\left[t(20) = -3.84, p < 0.01\right]$ or the masker (h_M) $\left[t(20) = -3.88, p < 0.01\right]$ stream compared to noise (h_N) $\left[t(20) = -1.85, p = 0.08\right]$or another story ($h_S$) $\left[t(20) = -0.85, p = 0.4\right]$. These results suggest that when participants failed to select the target keyword, or the masker keyword (if they mistakenly switched to the masker stream), they still listened to the presented speech streams. This could be because the failure to choose the target or masker keyword was caused by a failure to recall the correct word, rather than by a failure to attend the speech streams. Alternatively, it could be that, even when the listener's attention was directed elsewhere than to the target stream, primary automatic speech processes induced large-enough synchronous EEG activity to support reliable stimulus reconstruction accuracy. If so, using noise or an unrelated speech stimulus would necessarily lead to lower reconstruction accuracy. Further insight into this question may be gained by considering that no difference in reconstruction accuracy was noted, depending on whether extraneous keywords were replaced with targets (h_T) or with maskers (h_M) in challenging conditions. This result further suggests that, for these segments for which extraneous keywords were selected by participants, the participants were either dividing their attention across the two streams or listening to the mixture; this provides further justification for using the mixture as a filler (h_+). In addition, a difference in reconstruction was observed in the easiest listening condition when the extraneous keywords were filled with targets (h_T) or maskers (h_M): the reconstruction was improved with targets (h_T), to the point of matching a reconstruction performed with mixtures (h_+) as fillers. This finding suggests that it is reasonable to consider that participants have no (or nearly none) attentional switches when the difference between the target and masker voices is large enough.

The attentional scope extrapolation method, which was used to infer where the attention was directed in-between and around keywords, also influenced reconstruction accuracy. Inferred stimuli that modeled that the participant listens to the target even outside the keyword windows achieved a better reconstruction; this again suggests that it may be reasonable to assume that listeners are able to stay focused on the target throughout the trial.

Effect Size of the Enhancement

Several studies have previously shown that it is possible to improve reconstruction accuracy through different approaches. For example, properly choosing a regularization method and an adequate parameter for the decoders can lead to better stimulus reconstruction accuracy of 10–20% (Crosse et al., 2016; Wong et al., 2018). Similarly, Montoya-Martínez et al. (2021) showed that by optimizing the number of electrodes used for reconstruction, it was possible to improve a median score of 0.17–0.22, which represents a gain of 29%. The improvement observed in our results in the challenging condition enables to increase the reconstruction accuracy from 0.09 to 0.11, which represents a gain of 22%. Therefore, the approach we present here yields an improvement in reconstruction accuracy comparable to other techniques reported in the literature.

Acoustic Cues and Attention Switch Control

Spatial separation and voice differences are amongst the most important cues for auditory speech segregation. Teoh and Lalor (2019) improved auditory attention decoding accuracy by incorporating spatial attentional focus whereas Bednar and Lalor (2020) successfully reconstructed the spatial trajectory of a moving attended speech stream. The comparison of our results with the latter study is arduous due to methodological differences. Indeed, Bednar and Lalor's (2020) approach to reconstructing the spatial trajectory of a constantly moving attended speaker differs from our method in two ways. First, they directly manipulated the spatial location of the sources and this information is contained in the stimuli themselves. In contrast, in our experiment, the speakers' position was fixed and the attentional switches we captured with the behavioral responses were spurious rather than controlled. Second, Bednar and Lalor (2020) used a continuous variation of the location over time, whereas our behavioral account of attention is temporally restricted to three time windows corresponding to the three keyword positions. Between these keywords, we had to infer the participants' focus of attention. Finally, while spatial location translates into continuous angles, our behavioral information is ternary (target, masker, or extraneous). Therefore, transposing the method introduced by Bednar and Lalor (2020) to our behavioral account of attention does not seem straightforward. Yet, such an approach would deserve further investigation, perhaps combining it with potential acoustic and linguistic correlates of attentional switches (such as fluctuations in local target-to-masker ratio, or overlap in semantic context across target and masker).

Further Considerations and Conclusion

Results presented in the present study show that an enhancement of the stimuli reconstruction can be achieved in challenging situations where attention is modulated by voice cues such as F0 and VTL. By monitoring a participant's attentional focus, it is possible to obtain a better reconstruction of the real attended speech and therefore a better cortical representation. The advantage of parametric voice manipulation, as introduced in this article, is that the listening difficulty can be controlled. By generating extremely challenging conditions, it is possible to approach listening situations that share similarities with those experienced by people with hearing loss. For instance, CI users do not seem to efficiently benefit from voice cues, such as F0 and VTL, to discriminate two speech streams (Gaudrain and Başkent, 2018; El Boghdady et al., 2019). This was also the case for the participants of this present experiment, under challenging listening conditions. However, to further understand how voice-based speech segregation is hindered in listeners with hearing loss, more studies need to be conducted either with actual HI or CI listeners (e.g., Somers et al., 2018; Paul et al., 2020), or using hearing loss or electrical stimulation simulations, which can allow researchers to focus on specific aspects of sensory degradation.

As mentioned earlier, our approach is based on a temporally restricted measure of attention. In fact, results showing that the shortest attentional scope (i.e., target windows) works better than the longest scope (i.e., segments) underline the need for a more precise temporal resolution. One way to extend this temporal measurement would be to ask participants to press a button whenever they listen to a target stimuli within the target stream similarly to Crosse et al. (2015) with a hit/false-alarm/miss scoring. This approach would provide a better temporal resolution even though it introduces a dual task. Furthermore, this approach would allow for a greater comprehension of attentional bottom-up cues in speaker reconstruction and decoding studies.

One of the major challenges in neural tracking studies is to identify, based on brain activity, the speaker that the participant is listening to in a cocktail party situation. Our results stress the importance of incorporating attentional-switch tracking in speech enhancement or noise-reduction algorithms in hearing-aids.

DATA AVAILABILITY STATEMENT

The data used in this study are available here: https://doi.org/10.5281/zenodo.5680384.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

M-PH, CM, and EG designed the experiment. M-PH and EP performed data worked on collecting data and logistics. M-PH and EG analyzed the data. M-PH, CM, EP, and EG wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the LabEx CeLyA ("Centre Lyonnais d'Acoustique," ANR-10-LABX-0060) operated by the French National Research Agency.

ACKNOWLEDGMENTS

The authors thank Alexandra Corneyllie for her technical support; Fanny Meunier, Carolyn McGettigan, and Deniz Başkent for comments on earlier versions of this manuscript.

REFERENCES

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., and Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *Neuroimage* 124, 906–917. doi: 10.1016/j.neuroimage.2015.09.048

Akram, S., Simon, J. Z., and Babadi, B. (2017). Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments. *IEEE Trans. Biomed. Eng.* 64, 1896–1905. doi: 10.1109/TBME.2016.2628884

Bashford, J. A., Riener, K. R., and Warren, R. M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Percept. Psychophys.* 51, 211–217. doi: 10.3758/BF03212247

Başkent, D., and Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *J. Acoust. Soc. Am.* 139, EL51–EL56. doi: 10.1121/1.4942628

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv* [Preprint] ArXiv14065823,

Bednar, A., and Lalor, E. C. (2020). Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. *Neuroimage* 205:116283. doi: 10.1016/j.neuroimage.2019.116283

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 402–412. doi: 10.1109/TNSRE.2016.2571900

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.* 77, 1465–1487. doi: 10.3758/s13414-015-0882-9

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Crosse, M. J., Butler, J. S., and Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* 35, 14195–14204. doi: 10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* 10:604. doi: 10.3389/fnhum.2016.00604

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am.* 114, 2913–2922. doi: 10.1121/1.1616924

Ding, N., and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U. S. A.* 109, 11854–11859. doi: 10.1073/pnas.1205381109

El Boghdady, N., Gaudrain, E., and Başkent, D. (2019). Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users? *J. Acoust. Soc. Am.* 145, 417–439. doi: 10.1121/1.5087693

Enders, G., Enders, J., and Liber, I. (2015). *Le Charme Discret de L'intestin: Tout Sur un Organe Mal Aimé*. Paris: Éditions de Noyelles.

Festen, J. M., and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.* 88, 1725–1736. doi: 10.1121/1.400247

Gaudrain, E., and Başkent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear Hear.* 39, 226–237. doi: 10.1097/AUD.0000000000000480

Glasberg, B. R., and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-T

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Holtze, B., Jaeger, M., Debener, S., Adiloglu, K., and Mirkovic, B. (2021). Are they calling my name? Attention capture is reflected in the neural tracking of attended and ignored speech. *Front. Neurosci.* 15:643705. doi: 10.3389/fnins.2021.643705

Huet, M.-P. (2020). *Voice Mixology at a Cocktail Party: Combining Behavioural and Neural Tracking for Speech Segregation*. Ph.D. Thesis. Lyon: University of Lyon.

Huet, M.-P., Micheyl, C., Gaudrain, E., and Parizet, E. (2018). Who are you listening to? Towards a dynamic measure of auditory attention to speech-on-speech. *Interspeech* 2018, 2272–2275. doi: 10.21437/Interspeech.2018-2053

Huet, M.-P., Micheyl, C., Gaudrain, E., and Parizet, E. (2021). Vocal and semantic cues for the segregation of long concurrent speech stimuli in diotic and dichotic listening–the long-SWoRD test. *J. Acoust. Soc. Am.* 150. doi: 10.1121/10.0007225

Ives, D. T., Vestergaard, M. D., Kistler, D. J., and Patterson, R. D. (2010). Location and acoustic scale cues in concurrent speech recognition. *J. Acoust. Soc. Am.* 127, 3729–3737. doi: 10.1121/1.3377051

Jaeger, M., Mirkovic, B., Bleichner, M. G., and Debener, S. (2020). Decoding the attended speaker from EEG using adaptive evaluation intervals captures fluctuations in attentional listening. *Front. Neurosci.* 14:603. doi: 10.3389/fnins.2020.00603

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Commun.* 27, 187–207. doi: 10.1016/S0167-6393(98)00085-5

Mathôt, S., Schreij, D., and Theeuwes, J. (2012). OpenSesame: an open-source, graphical experiment builder for the social sciences. *Behav. Res. Methods* 44, 314–324. doi: 10.3758/s13428-011-0168-7

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Middlebrooks, J. C. (2017). "Spatial stream segregation," in *The Auditory System at the Cocktail Party. Springer Handbook of Auditory Research*, eds J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Cham: Springer), 137–168. doi: 10.1007/978-3-319-51662-2_6

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., and Babadi, B. (2018). Real-time tracking of selective auditory attention from M/EEG: a bayesian filtering approach. *Front. Neurosci.* 12:262. doi: 10.3389/fnins.2018.00262

Montoya-Martínez, J., Vanthornhout, J., Bertrand, A., and Francart, T. (2021). Effect of number and placement of EEG electrodes on measurement of neural tracking of speech. *PLoS One* 16:e0246769. doi: 10.1371/journal.pone.0246769

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional Selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Paul, B. T., Uzelac, M., Chan, E., and Dimitrijevic, A. (2020). Poor early cortical differentiation of speech predicts perceptual difficulties of severely hearing-impaired listeners in multi-talker environments. *Sci. Rep.* 10:6141. doi: 10.1038/s41598-020-63103-7

Pyschny, V., Landwehr, M., Hahn, M., Walger, M., von Wedel, H., and Meister, H. (2011). Bimodal hearing and speech perception with a competing talker. *J. Speech Lang. Hear. Res.* 54, 1400–1415. doi: 10.1044/1092-4388(2011/10-0210)

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Singmann, H., Bolker, B., Westfall, J., and Aust, F. (2019). *Afex: Analysis of Factorial Experiments*.

Somers, B., Verschueren, E., and Francart, T. (2018). Neural tracking of the speech envelope in cochlear implant users. *J. Neural Eng.* 16:016003. doi: 10.1088/1741-2552/aae6b9

Søndergaard, P. L., and Majdak, P. (2013). "The auditory modeling toolbox," in *The Technology of Binaural Listening*, ed. J. Blauert (Heidelberg: Springer), 33–56. doi: 10.1007/978-3-642-37762-4_2

Søndergaard, P. L., Torrésani, B., and Balazs, P. (2012). The linear time frequency analysis toolbox. *Int. J. Wavelets Multiresolution Inf. Process.* 10:1250032. doi: 10.1142/S0219691312500324

Teoh, E. S., and Lalor, E. C. (2019). EEG decoding of the target speaker in a cocktail party scenario: considerations regarding dynamic switching of talker location. *J. Neural Eng.* 16:036017. doi: 10.1088/1741-2552/ab0cf1

Vestergaard, M. D., Ives, D. T., and Patterson, R. D. (2009). The advantage of spatial and vocal characteristics in the recognition of competing speech. *Proc. Int. Symp. Audit. Audiol. Res.* 2, 535–544.

Wong, D. D. E., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., and de Cheveigné, A. (2018). A comparison of regularization methods in forward and backward models for auditory attention decoding. *Front. Neurosci.* 12:531. doi: 10.3389/fnins.2018.00531

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R, Statistics for Biology and Health*. New York, NY: Springer, doi: 10.1007/978-0-387-87458-6

Conflict of Interest: CM was employed by the company Starkey France, S.a.r.l.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Huet, Micheyl, Parizet and Gaudrain. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.