Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field Environment

Patrick J. Schäfer¹ ², Farah I. Corona-Strauss¹ ², Ronny Hannemann³, Steven A. Hillyard⁴ ⁵, and Daniel J. Strauss¹ ²

Abstract
Auditory attention can be defined as the cognitive process that enables us to selectively focus on relevant aspects of the acoustic environment while other aspects are ignored. The remarkable ability of the auditory system to focus on one out of several speakers in a multispeaker environment has become known as the *cocktail party effect*. Although the neural processes underlying selective auditory attention (SAA) are not well understood, it has recently been shown that the cortical representation of a listener’s attended sound stream can be recorded noninvasively from the scalp and that stimulus reconstruction from single trial electroencephalographic (EEG) data enables the decoding of the orientation of auditory attention. The present study extends this approach by evaluating its efficacy in a naturalistic and challenging four-speaker acoustic free field environment, in which the four speakers were spatially separated and presented different but equally salient spoken messages to the listeners. Ten participants were instructed to focus SAA on a spoken prose message in one of the four loudspeakers while ignoring the remaining three streams of prose. Concurrent EEG activity recorded via 128 scalp channels was used for a stimulus reconstruction analysis. The results showed that this approach can be used to decode the orientation of SAA even in a complex and realistic acoustic setting. To confirm that the successful decoding was driven by correspondences between the recorded EEG activity and the attended speech envelopes, the analysis method was validated against randomly constructed sets of surrogate data and by correlations with behavioral data.

Keywords
selective attention, auditory attention, cocktail party problem, stimulus reconstruction, electroencephalogram

Date received: 21 February 2018; revised: 23 October 2018; accepted: 29 October 2018

Introduction
In 1953, E. C. Cherry introduced the term *cocktail party effect* to describe a listener’s ability to attend easily to one out of several competing and spatially separated talkers in a multitalker environment such as a noisy cocktail party (Cherry, 1953). The cognitive process that enables us to selectively focus on relevant information in our sensory environment while ignoring or suppressing concurrent irrelevant information is known as attention (Treisman, 1969). Solving the cocktail party problem involves using selective auditory attention (SAA) to focus on one of several competing sound sources. Among other things, this typically involves the auditory system’s ability to spatially direct auditory attention.

SAA relies on neural and cognitive processes that are not fully understood. However, multiple studies have revealed how specific regions of the cortex interact during auditory scene analysis and have described

¹Systems Neuroscience and Neurotechnology Unit, Neurocenter, Faculty of Medicine, Saarland University, Homburg/Saar, Germany
²Key Numerics GmbH, Saarbrücken, Germany
³Sivantos GmbH, Erlangen, Germany
⁴Leibniz-Institute for Neurobiology, Magdeburg, Germany
⁵Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA

Corresponding Author:
Daniel J. Strauss, Systems Neuroscience & Neurotechnology Unit, Saarland University Hospital, Building 90.5, D–66421 Homburg/Saar, Germany. Email: daniel.strauss@uni-saarland.de

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
The present study extends previous work by investigating the effectiveness of the stimulus reconstruction approach in a naturalistic and challenging four-speaker acoustic free field environment. In particular, this is the first study in which four speakers presenting different but equally salient spoken messages have been used to investigate the classification accuracy of the stimulus reconstruction approach. A previous study by Fuglsang, Dau, and Hjortkjaer (2017) employed eight spatially separated loudspeakers, but six of them played background noise and only two served as target loudspeakers (Fuglsang et al., 2017). In the present design, the loudspeakers were equal in relation to the task of the participant (each of the four loudspeakers served as the target speaker during the experiment), equal in relation to the presented content (each loudspeaker played its own continuous audio book), and equal in volume (all loudspeakers were identical in volume due to careful calibration). Moreover, this is the first study in which subjects were confronted with more than two spatially separated speakers of equal relevance in a free-field environment. The previous study of O’Sullivan et al. (2017) presented subjects with more than two target speakers, but all speakers were presented via one loudspeaker (O’Sullivan et al., 2017). Here, particular emphasis was placed on positioning the four loudspeakers at locations relevant to future hearing aid applications; that is, the frontal half circle—the area that is essentially covered by the microphones of hearing aids. In addition, the stimulus material was selected to offer complex as well as naturalistic sentences, that is, longer passages of classic fiction, as compared to the simple presentation of digits or short nonsense sentences as, for example, in the Oldenburger Sentence Test. The intention here was to evaluate the stimulus reconstruction approach in an environment that was as close to a real-life situation as possible, that is, in an environment having many different but relevant spatially separated speakers in a listener’s entire field of vision—as might be encountered in a cocktail party.

In sum, the present study aimed to explore the limits of the stimulus reconstruction approach in a complex and life-like acoustic environment with several spatially separated talkers of equivalent relevance. To verify that the classification accuracy of the stimulus reconstruction approach was driven by regularities within the EEG, we tested the approach against two randomly constructed sets of surrogate data and compared the classification accuracy with a set of behavioral data. The long-range goal of this research is to assess the possibility of moving stimulus reconstruction from laboratory settings into everyday situations, including the future development of advanced hearing aids.
Materials and Methods

Participants

In total, 10 subjects with no known health problems—especially no known hearing loss (tested with a pure tone audiometer) or neurophysiological diseases/impairments—took part in this study. The participants’ age ranged between 22 and 27 years (average: 24.5 ± 1.9 years). Eight of the subjects were male and nine were right handed. Every participant was a German native speaker. Immediately before the measurement, the subjects were informed about the procedure and the objectives of the measurement and gave their written consent. The design of the experiment was planned in accordance with ethics guidelines and the Declaration of Helsinki, and the study was approved by the local ethics committee (Arztekammer des Saarlandes—Medical Council of the Saarland, Germany). Every subject had the free choice to abandon the procedure and withdraw their participation at any time.

Experimental Setup

A schematic of the acoustic environment used for this study can be seen in Figure 1. To achieve an acoustically controlled environment—especially to achieve as little sound reflection as possible—a cubic, acoustically controlled room was used. The dimensions of the room were 3 × 3 × 3 m. The walls and the ceiling were equipped with heavy (900 g/m²) stage molton, that is, a curtain reducing sound reflections. Within the room, a circular arranged acoustic free field system with four active loudspeakers (Neumann KH 120 A, Georg Neumann GmbH, Germany) and a diameter of 2 m was implemented. The speakers were equidistantly arranged in a half circle—one speaker at −90°, one at −30°, one at +30°, and one +90° (LS1, LS2, LS3, and LS4, respectively)—at the height of the subject’s ear. The loudspeakers were controlled by a USB 2.0 audio interface (Scarlett 18i20, Focusrite, USA). A comfortable chair and a table with a monitor on it were placed in the middle of the free field system. In addition, a chin rest was installed exactly in the center. At the end of the table at 0° at eye level, a computer monitor was used to present visual feedback and relevant instructions. To record the subject’s responses, a computer keyboard was placed on the table (Figure 1).

Paradigm and Stimulus Material

The paradigm was designed to simulate the cocktail party problem. The participant’s goal was to follow the spoken content of an audiobook presented by one of the four loudspeakers, while ignoring the remaining three

Figure 1. Setup of the stimulus presentation and the data acquisition: PC 1 controls the paradigm. It provides the stimulus material, that is, four different audiobooks and the trigger signal, to the audio interface. The audio interface deploys the stimulus material to the four loudspeakers (LS1–LS4) and the trigger signal to the trigger box. In addition, PC 1 is connected to the monitor and the keyboard in front of the participant—it broadcasts the instructions on the screen and receives the input from the keyboard. PC 2 handles the data acquisition. It records the data from the signal amplifier, that is, 128 channels of EEG data from scalp electrodes and the trigger signal.
loudspeakers that each played a different audiobook at the same time. Four different audiobooks spoken by four different professional speakers (two male and two female voices) were used to generate the acoustic environment. Each of the audio books was professionally recorded and sampled with a frequency of 44.1 kHz (Ackner & Fischbach, 2017).

- LS1 (−90°) In the Penalty Colony, F. Kafka—Spoken by a man (Johannes Gabriel)
- LS2 (−30°) The Earthquake in Chile, H. von Kleist—Spoken by a woman (Brigitte Truebenbach)
- LS3 (+30°) The Stone Heart, E.T.A. Hoffmann—Spoken by a man (Thomas Dahler)
- LS4 (+90°) The Young King, O. Wilde—Spoken by a woman (Maja Chrenko)

Each audiobook was decomposed into 25 segments, each having a duration of 120 s. Throughout the experiment, each segment was presented in chronological order to maintain the story line. Each of the 100 segments was calibrated individually to 55 dBAeq. A hand-held sound-level meter (type 2250, Brüel & Kjær, Denmark) was used to calibrate the audiobook segments at the position of the ear. The signal to noise ratio between the different loudspeakers was thus held constant at 0 dB over the whole experiment. To control the entire paradigm, the open source toolbox Psychophysics Toolbox Version 3 (PTB-3) for Matlab (MATLAB R2013A, MathWorks, USA) was used (Brainard, 1997; Pelli, 1997).

**Experimental Design**

The experiment was divided into three parts—a training session, experiment 1 (E1), and experiment 2 (E2). The training session consisted of one trial. The participant was asked to sit as comfortable as possible, to rest his or her chin on the chin rest, and to listen to the audiobook presented at LS3 for 120 s while ignoring the audiobooks presented from the other three loudspeakers (LS1, LS2, and LS4). The participant was also asked to move as little as possible and to fixate his or her eyes upon the cross presented on the monitor during the presentation of the audiobooks. Shortly before the acoustic stimuli were played, three content-related questions were displayed on the screen. The subject was told that it was not necessary to keep the questions in mind because the questions were going to be displayed again after the trial with multiple choice response options. The reason for displaying the questions before the stimulus presentation was to help the participant to focus on the corresponding audiobook segment. Immediately after presentation of the audiobook segment, the subject had to answer the three previewed multiple choice questions displayed on the screen by using the keyboard. The subject was asked if the training trial was clear or if there were any further questions. If there were none, the participant was told that the following trials were going to be exactly like the training trial and that it was up to him or her to control the speed of the experimental procedure, and that it was possible to take rests.

E1 consisted of eight trials (120-s segments). Each of the four loudspeakers played its individual audiobook segment, while the subject was asked to focus either on the far left or on the far right speaker, that is, LS1 or LS4, respectively. The participant was asked to focus to LS1 or to LS4 for four of the trials each—the exact order was randomized for every subject. E2 consisted of 16 trials, and the participant had to pay attention to each of the four loudspeakers for four trials each. Here, the exact order was randomized for each participant as well.

**Data Acquisition**

EEG was recorded from 128 active Ag/AgCl electrodes and a ground electrode. Of these, 32 EEG channels were used in the stimulus reconstruction analysis. The electrodes had been positioned using an EEG-cap (Active Electrode System, g.GAMMAsys, Guger Technologies, Austria) that followed the International 10-20-System. The impedance of each electrode was kept below 50 kΩ as recommended for active electrodes. Scalp EEG activity was amplified (g.Hiamp, Guger Technologies, Austria) and sampled at 512 Hz with reference to a ground electrode positioned at the center of the forehead (AFz). No additional online processing options such as frequency filtering were used. The signal amplifier was connected to the acquisition PC (PC 2) using USB 2.0. To control the data acquisition, a Simulink interface (MATLAB R2013A, MathWorks, USA) was used. A trigger signal indicated the onset of each experimental trial. The trigger signal was deployed to the signal amplifier via paradigm PC (PC 1), sound card, and conditioner box (g.TRIGbox, Guger Technologies, Austria). Thus, the EEG data could be analyzed in synchrony with the presentation of the target sound. In addition, PC 1 was used to record the answers to the multiple choice questions presented to the subject.

**Data Processing**

The acquired EEG (raw) data were imported into the software Matlab (MATLAB R2017A, MathWorks, USA). The data were stored as a \( N \times T \) matrix, where \( N \) denotes the number of recorded EEG channels and \( T \) denotes the recorded data points. The first step of data processing was to rereference the recorded channels against the electrode positioned at the vertex (Cz). The next step was to band pass filter all channels (from 1 Hz to 45 Hz) using a zero phase shift finite impulse response
(FIR) band pass filter of order 1000 based on a Hamming window. The filtered EEG channels were segmented into 24 blocks according to the recorded trigger signal, that is, 8 for E1 and 16 for E2. Each of those data matrices represented the preprocessed EEG data of one experimental trial: \( r_K(t, n) \), \( K = 1, \ldots, 24, \ t = 1, \ldots, 61, 440 \) and \( n = 1, \ldots, 128 \). To ensure that the data quality was sufficient, each EEG channel in each matrix was transformed into the frequency domain using Fast Fourier Transformation and was visualized afterwards. The resulting spectrograms were individually checked for EEG-typical patterns, that is, peaks occurring between 8 Hz and 13 Hz resulting from \( \alpha \)-activity. If the described peak was detectable, the data set was categorized as analyzable. The peak was defined as detectable if the ratio between the power of the area \( \pm 2 \) Hz around the \( \alpha \)-peak and the power of the surrounding 2 Hz was greater than one. An example of this spectral analysis is shown in Figure 2. In the figure, the peaks in the ranges of the \( \alpha \)- and the \( \beta \)-bands, that is, in the ranges of 8 to 13 Hz and 13 to 25 Hz, respectively, can be clearly recognized which indicates that the quality of the recorded EEG data was sufficiently good.

The data sets that had been categorized as analyzable were prepared for further application of the stimulus reconstruction algorithm, following the guidelines described by Crosse et al. (2016). At first, the data volume was reduced. Of the recorded 128 channels, only 32 were included into the analysis. Those 32 channels were equally distributed across the scalp according to the International 10-20-System. The next step was to filter the 32 EEG channels using a zero-phase shift FIR low-pass filter with a cut-off frequency at 15 Hz and order 1000 based on a Hamming window. In addition, the amplitudes of every EEG channel were individually scaled to 0 to 1. The EEG data sets were then downsampled to 128 Hz and segmented into subblocks of 30 s duration. This resulted in 96 matrices per participant: \( r_L(t, n) \) with \( L = 1, \ldots, 96, \ t = 1, \ldots, 3840 \) and \( n = 1, \ldots, 32 \).

The stimulus reconstruction algorithm requires the recorded EEG data together with the physical characteristics of the acoustic stimulation as input. Thus, the stimulus material had to be prepared in the same way as the EEG data, that is, having the same duration and the same sampling frequency. The first step was to calculate the broadband envelope of each audiobook segment according to the following equation:

\[
x_a(t) = x(t) + i \cdot \hat{x}(t)
\]

where \( x_a(t) \) represents the complex analytic signal resulting from the complex sum of the audiobook segment \( x(t) \) and its representation in Hilbert space \( \hat{x}(t) \). The speech envelope is defined as the absolute value of \( x_a(t) \), that is, \( e(t) = |x_a(t)| \). The envelope was sampled down from 44.1 kHz to 128 Hz following application of an anti-aliasing filter. The next step was to normalize each envelope between 0 and 1. The final step was to separate the calculated envelopes into 30-s snippets so as to match their corresponding EEG segments: \( e_L(t) \) with \( L = 1, \ldots, 24 \) and \( t = 1, \ldots, 3840 \).

---

**Figure 2.** Normalized spectrogram of the recorded EEG data. The example shows subject 3 while focusing on speaker LS1 for the first time in E1. On the x-axis, the frequency is noted in Hz and on the y-axis, the magnitude. Each line represents one of the 32 EEG channels. One can clearly recognize peaks in the range of \( \alpha \)- (8–13 Hz) and in the range of \( \beta \)- (13–25 Hz) activity, which is a typical pattern for EEG data.
**Data Analysis**

The following gives a short overview of the stimulus reconstruction approach. We refer to Crosse et al. (2016) for more detailed information.

The basic idea behind the approach is that the cortex acts like a linear time invariant system mapping input, that is, acoustic stimuli, to a certain output, that is, the EEG activity. ERPs in the EEG can be interpreted as the impulse response of the cortex to a discrete stimulation—for example, a click sound. Following that idea, the ongoing EEG activity resulting from ongoing, continuous stimulation—like real speech—can be interpreted as a linear convolution. According to this, the instantaneous neural activity \( r(t, n) \) \((t = 1 \ldots T\) denotes points in time and \( n = 1 \ldots N\) denotes EEG channels\) is the result of a convolution of the acoustic stimulation \( s(t)\), that is, the speech envelope, with an unknown, channel-specific temporal response function (TRF) \( w(\tau, n)\). The TRF can be seen as a filter describing the transformation of the ongoing stimulus to the ongoing EEG activity resulting from ongoing, continuous stimulation. ERPs in the EEG can be interpreted as the impulse response of the cortex to a stimulus feature. Those time lags result from the cross-validation process, one decoder was calculated using \( N - 1 = 96 - 1 = 95\) data sets and the corresponding 95 speech envelopes for every given ridge parameter \( \lambda\).

The calculated 95 \( \times \) 11 decoders were averaged along the 95 trials to prevent over-fitting. The results were 11 averaged decoders \( \tilde{g}_L(\tau, n)\), that is, one for every ridge parameter. These were used to calculate the reconstructed speech envelopes. The 11 reconstructed speech envelopes \( \hat{e}_L(t)\) were compared to the actual speech envelope \( e(t)\) using Pearson’s correlation coefficient. The described procedure was repeated 96 times—one time for every trial. As a result, we obtained a 96 \( \times \) 11 matrix (96 trials and 11 ridge parameters) with reconstructed speech envelopes \( \hat{e}_{L,\lambda}(t)\), as well as a 96 \( \times \) 11 matrix with correlation coefficients \( \rho_{L,\lambda}\). The average along the 96 trials was calculated for the correlation coefficients, and their maximum was determined.

Here, \( \hat{e}_L(t)\) represents the optimal set of averaged decoders for every subject, that is, \( \tilde{g}_{L,\lambda}(\tau, n)\), and to compute the set of optimal reconstructed speech envelopes \( \hat{e}_{L,\lambda}(t)\).

The second step of the data analysis involved the comparison of the reconstructed speech envelope \( \hat{e}_{L,\lambda}(t)\) with the corresponding attended speech envelope \( e_{\lambda}(t)\) and with the distractor envelopes \( a_L(t), b_L(t), \) and \( c_L(t)\). This was done by calculating Pearson’s correlation coefficients between \( \hat{e}_{L,\lambda}(t)\) and \( e_{\lambda}(t), a_{L}(t), b_{L}(t), \) and \( c_{L}(t)\), respectively. The result was a 96 \( \times \) 4 matrix

\[
R = \begin{bmatrix}
    r(1, 1) & \ldots & r(1, 1) & 0 \\
    \vdots & \ddots & \vdots & \vdots \\
    r(T, 1) & \ldots & r(T, 1) & r(T - \tau_M, 1) \\
    0 & r(T, 1) & \ldots & r(T, 1)
\end{bmatrix}
\]

Here, \( I \) represents the identity matrix and \( \lambda \) is a smoothing constant or ridge parameter. The ridge parameter can be adjusted using a leave-one-out cross-validation process to maximize the correlation between \( s(t) \) and \( \hat{s}(t)\).

The first step of data analysis was the reconstruction of the attended speech envelopes \( e_{\lambda}(t)\) using the 96 prepared EEG data sets \( r_L(t, n)\) for every subject. For this, the previously mentioned decoder \( g \) had to be computed and the ridge parameter \( \lambda \) also had to be optimized. This was achieved by using a leave-one-out cross-validation process. In our case, the time lag range was defined as \( \tau_{\text{min}} = -50 \text{ ms} \) and \( \tau_{\text{max}} = 350 \text{ ms} \), and the following ridge parameters were tested: \( \lambda = 2^0, 2^2, \ldots, 2^20 \). During the cross-validation process, one decoder was calculated using \( N - 1 = 96 - 1 = 95 \) data sets and the corresponding 95 speech envelopes for every given ridge parameter \( \lambda \).

The calculated 95 \( \times \) 11 decoders were averaged along the 95 trials to prevent over-fitting. The results were 11 averaged decoders \( \tilde{g}_L(\tau, n)\), that is, one for every ridge parameter. These were used to calculate the reconstructed speech envelopes. The 11 reconstructed speech envelopes \( \hat{e}_L(t)\) were compared to the actual speech envelope \( e(t)\) using Pearson’s correlation coefficient. The described procedure was repeated 96 times—one time for every trial. As a result, we obtained a 96 \( \times \) 11 matrix (96 trials and 11 ridge parameters) with reconstructed speech envelopes \( \hat{e}_{L,\lambda}(t)\), as well as a 96 \( \times \) 11 matrix with correlation coefficients \( \rho_{L,\lambda}\). The average along the 96 trials was calculated for the correlation coefficients, and their maximum was determined. Where the maximum value was reached, the optimum of the ridge parameter \( \lambda \) was found. Now, it was possible to define the optimal set of averaged decoders for every subject, that is, \( \tilde{g}_{L,\lambda}(\tau, n)\), and to compute the set of optimal reconstructed speech envelopes \( \hat{e}_{L,\lambda}(t)\).

The second step of the data analysis involved the comparison of the reconstructed speech envelope \( \hat{e}_{L,\lambda}(t)\) with the corresponding attended speech envelope \( e_{\lambda}(t)\) and with the distractor envelopes \( a_{L}(t), b_{L}(t), \) and \( c_{L}(t)\). This was done by calculating Pearson’s correlation coefficients between \( \hat{e}_{L,\lambda}(t)\) and \( e_{\lambda}(t), a_{L}(t), b_{L}(t), \) and \( c_{L}(t)\), respectively. The result was a 96 \( \times \) 4 matrix...
containing the resulting correlation coefficients $\rho_{L, SE}$, where $SE = 1, \ldots, 4$ denotes the different speech envelopes. For every trial, it was necessary to determine where the correlation coefficients reached their maximum. The position where the correlation coefficient was maximal defined the attended classified speech envelope. If the classified attended speech envelope matched the actual attended speech envelope, the classification was classified as correct; otherwise, it was classified as incorrect. The final result was the total number of correct classifications, which could range between 0 and 96 for each participant. Given that there were four different loudspeakers, we considered the classification of the orientation of SAA to be successful if the total number of correct classifications reached a value well above the chance level, that is, $0.25 \cdot 96 = 24$. If the loudspeaker to which a subject directs his or her attention is correctly classified as such at least 36 (out of 96) times, then the null hypothesis that the data are uniformly distributed can be rejected with a significance level of $p \leq 0.05$ using the $\chi^2$ test. The classification accuracy of the decoder is then assumed to be significantly higher than that which would be achieved by simple guessing, that is, well above the chance level. The foregoing data analysis procedure is visualized in Figures 3 through 6.

To validate this stimulus reconstruction method, the abovementioned analysis was carried out on two sets of randomized surrogate data. The first set was generated by combining the EEG data from each trial with attended speech envelopes from randomly chosen trials. The second set was generated by randomly combining the EEG data of each trial with a temporally matching speech envelope that could be either attended (25%) or unattended (75%). If our assumptions are correct, that is, that SAA specifically enhances the neural representation of the attended speech message, the total number of correct classifications for these surrogate data should be in the range of simple guessing, that is, a value around 24.

**Results**

After every experimental trial, the participants had to answer three content-related multiple choice questions with four possible answers so that chance level was 25%. The aim was to verify that the participants were able to follow the story line of the audiobook segments they were asked to attend to. Figure 7 shows the percentage of correctly answered multiple choice questions for each subject, which ranged between 50.0% and 94.4% with an average of $76.8 \pm 13.1\%$.

To find out whether it was possible to predict the orientation of SAA from the combination of recorded neural activity and the physical characteristics of the spoken messages, that is, the speech envelopes, in a challenging *cocktail party situation*, we used stimulus reconstruction and counted the number of correct and incorrect classifications. Figure 8 shows the results for every participant over each of the 24 experimental trials. Note that each trial was subdivided into four parts, resulting in 96 segments to be classified. The number of correct classifications per subject ranged between 21 and 81 with an average of $58.70 \pm 18.02$. $\chi^2$ test carried out for each subject showed these classifications to be highly significant: With exception of subject 2, significance levels in each case were $p < 10^{-5}$ or

---

**Figure 3.** The $N - 1$ preprocessed EEG data, combined with the corresponding attended speech envelopes, were used to calculate a decoder for each of the $N - 1$ trials and each ridge parameter. The computed decoders were averaged along the trials in order to prevent over-fitting.
better when tested against the null hypothesis that the classifications come from a uniform distribution. The significance level over all subjects (with 2 excluded) was $p = 3.9 \times 10^{-7} \pm 1.2 \times 10^{-6}$. For subject 2, the classifications did not differ significantly from chance $p = .68$.

To demonstrate the validity of these classifications, the stimulus reconstruction algorithm was also applied to two sets of randomly constructed surrogate data. For the first set of surrogate data, the recorded EEG activity from each trial was matched with attended speech envelopes of randomly selected trials. In this case, the total number of correct classifications (of the attended speaker) fell within the range of 16 and 32 per subject with an average of $25.40 \pm 4.67$. For the second set of surrogate data, the EEG activity was combined with the
speech envelopes of matching trials, but not necessarily the attended envelopes. Here, the total number of correct classifications (of the attended speaker) fell within the range of 22 and 36 per subject with an average of 28.30 ± 4.64.

To verify that the EEG-based classification accuracy was actually related to SAA, the number of correct classifications was correlated (using Pearson’s coefficient) with the number of correctly answered multiple choice questions on the content of the audiobooks. The correlation was positive (ρ = .69) and significant (p = .026).

**Discussion and Future Work**

The present results show that it is possible to decode the orientation of SAA in a four-speaker free field environment using stimulus reconstruction. Over all subjects, the EEG-based stimulus reconstruction algorithm correctly classified the attended loudspeaker on an average of 61.1% of the trials, well above chance level of 25%. The two best subjects showed over 80% correct classifications. These findings provide strong support for the hypothesis that the recorded EEG is mostly driven by the attended speaker compared to the unattended ones; in other words, the EEG is mostly driven by the rhythm of speech that the listener is attending. Figure 8 indicates that the classification was successful for every participant, except subject 2. A recent study also tested the stimulus reconstruction approach with several loudspeakers in free field (Fuglsang et al., 2017), but only two of the loudspeakers were used as targets of attention, while the others served as distractors. In that study, an
average classification accuracy rate of 87.1% was achieved, some 37.1 percentage points above the chance level of 50%. Similarly, the present study achieved a classification accuracy of 36.1% percentage points above the chance level of 25% using four spatially separated loudspeakers having equivalent relevance to the subjects. The present results thus demonstrate for the first time that the stimulus reconstruction approach can be used effectively to determine the direction of SAA in an environment having four spatially separated loudspeakers with equal relevance to the listener.

On average, the 10 subjects were able to correctly answer 76.8% of the multiple choice questions about the content of the attended audiobook story. This indicates that the subjects were able to attend effectively to the designated spoken message and that the present paradigm is a valid simulation of the cocktail party problem. Subject 2 stands out as the participant showing the weakest performance, with less than 50% correct answers. This subject also showed the lowest number of correct classifications of the attended loudspeaker based on EEG stimulus reconstruction (see Figure 8), which in fact did not exceed the chance level. It appears that subject 2 was unique in not focusing attention effectively on the attended message.

To validate this classification process based on EEG stimulus reconstruction, two sets of surrogate data were generated. The aim of the first set of surrogate data was to verify that classification of the attended and unattended speaker would not be possible if the algorithm was fed with obviously incorrect data. In Figure 9, it can be seen that the classification was indeed not successful, that is, not significantly above chance level. The goal of the second set of surrogate data was to determine whether classification was possible if the algorithm is randomly fed with attended and unattended speech envelopes from the same experimental trial. In Figure 10, it can be seen that the classification was not successful although its performance appeared to be slightly better than with the first set of surrogate data. This slight improvement with the second set of surrogate data might be explained as follows: The probability of combining a certain trial of EEG data with the matching attended speech envelope was $\frac{1}{96}$ for the first set of surrogate data, while the probability of combining a certain trial of EEG data with the actual attended speech envelope was $\frac{1}{4}$ for the second set. Nevertheless, the negative results obtained using the two sets of surrogate data indicate that the success of the classification process is based on the correspondence between the recorded neural activity and the speech envelopes and was not caused by an artifact in the algorithm itself.

Another noteworthy finding of the present study was that the number of correctly answered content questions correlated significantly (and positively) with the number of correctly classified trials. Assuming that the number of correctly answered questions depends on the listener’s attention to the designated audiobook, this would be another indication that the stimulus reconstruction approach is indeed sensitive to SAA. Based on this result, we can envisage future testing of the approach as an objective EEG-based measure of focused
attention and speech intelligibility in realistic acoustic environments.

A major aim of this study was to investigate the limitations of the stimulus reconstruction approach with respect to its potential applications for improving hearing aid capabilities. For this reason, we placed the speaker locations so as to cover the frontal semicircle in the free field environment, which corresponds to the zone of hearing aid microphones. In addition, we set up a realistic and complex listening environment consisting of four equally relevant and spatially separated speakers. This type of acoustic environment is particularly relevant for the further development of hearing aids, given that people with hearing loss have greater difficulty in understanding a speaker in a multispeaker environment compared to normal hearing individuals (e.g., Bernarding, Strauss, Hannemann, Seidler, & Corona-Strauss, 2013; Pichora-Fuller & Singh, 2006). For this reason, the stimulus reconstruction approach may have important applications in the field of rehabilitation audiology: The inclusion of a hearing aid wearer’s intended targets of attention into the design of future hearing aids, that is, by means of a brain computer interface, could help to improve the quality of life of hearing-impaired people.

Figure 9. The number of classifications of the attended and unattended speakers for each subject using the first set of surrogate data.

Figure 10. The number of classifications of the attended and unattended speakers for each subject using the second set of surrogate data.
An important advantage of the stimulus reconstruction approach in this regard is that it is based on single-trial EEG recordings and thus can provide near real-time information to the listener. The present findings suggest that the approach is robust enough to begin testing in everyday situations by using wearable EEG devices and new electrodes, such as in-ear electrodes, to help prepare the way for a new generation of hearing aids.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the German Federal Ministry of Education and Research grant (number BMBF-FZ 03FH004IX5).

ORCID iD
Patrick J. Schäfer https://orcid.org/0000-0002-3393-5967

References
Ackner, J., & Fischbach, D. (2017). vorleser.net. Buchfunk Verlag.
Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., & Corona-Strauss, F. I. (2013). Neural correlates of listening effort related factors: Influence of age and hearing impairment. Brain Research Bulletin, 91, 21–30. doi: 10.1016/j.brainresbull.2012.11.005.
Bidet-Caulet, A., Fischer, C., Besle, J., Aguer, P.-E., Giard, M.-H., & Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. Journal of Neuroscience, 27(5), 9252–9261.
Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society, 25(5), 402–412.
Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. Nature Reviews. Neuroscience, 14, 693–707. doi: 10.1038/nrn3565.
Brainard, D. H. (1997). The psychophysics toolbox. Spatial Vision, 10(4), 433–436. doi: 10.1163/156856897X00357.
Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America, 25, 974–979.
Crosse, M. J., DiLiberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. Frontiers in Human Neuroscience, 10, 604. doi: 10.3389/fnhum.2016.00604.
Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—Focusing the searchlight on sound. Current Opinion in Neurobiology, 17(4), 437–455.
Fuglsang, S. A., Dau, T., & Hjortkjaer, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. NeuroImage, 156, 435–444. doi: 10.1016/j.neuroimage.2017.04.026.
Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. Science, 182(4108), 177–180.
Lopez, M.-A., Pomares, H., Pelayo, F., Urquiza, J., & Perez, J. (2009). Evidences of cognitive effects over auditory steady-state responses by means of artificial neural networks and its use in brain-computer interfaces. Neurocomputing, 72, 3617–3623. doi: 10.1016/j.neucom.2009.04.021.
Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485, 233–236. doi: 10.1038/nature11020.
O’Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G.,… Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex (New York, N.Y.: 1991), 25(7), 1697–1706. doi: 10.1093/cercor/bht355.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10(4), 437–442.
Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiologic rehabilitation. Trends in Amplification, 10, 29–59. doi: 10.1177/108471380601000103.
Rieke, F., Bodnar, D. A., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Proceedings. Biological Sciences, 262, 259–265. doi: 10.1098/rspb.1995.0204.
Stanley, G. B., Li, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. Journal of Neuroscience, 19(18), 8036–8042.
Treisman, A. M. (1969). Strategies and models of selective attention. Psychological Review, 76(3), 282–299. doi: 10.1037/h0027242.
Wöstmann, M., Herrmann, B., Maess, B., & Olschewski, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. Proceedings of the National Academy of Sciences of the United States of America, 113(14), 3873–3878. doi: 10.1073/pnas.1523357113.
Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Scheyon, C. A., McKhann, G. M.,… Simon, J. Z. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron, 77, 980–991. doi: 10.1016/j.neuron.2012.12.037.