Cortical Auditory Attention Decoding During Music and Speech Listening

Adéle Simon, Gérard Loquet, Jan Østergaard, *Senior Member, IEEE*, and Søren Bech

***Abstract*— It has been demonstrated that from cortical recordings, it is possible to detect which speaker a person is attending in a cocktail party scenario. The stimulus reconstruction approach, based on linear regression, has been shown to be useable to reconstruct an approximation of the envelopes of the sounds attended to and not attended to by a listener from the electroencephalogram data (EEG). Comparing the reconstructed envelopes with the envelopes of the stimuli, a higher correlation between the envelopes of the attended sound is observed. Most of the studies focused on speech listening, and only a few studies investigated the performances and the mechanisms of auditory attention decoding during music listening. In the present study, auditory attention detection (AAD) techniques that have been proven successful for speech listening were applied to a situation where the listener is actively listening to music concomitant with a distracting sound. Results show that AAD can be successful for both speech and music listening while showing differences in the reconstruction accuracy. The results of this study also highlighted the importance of the training data used in the construction of the model. This study is a first attempt to decode auditory attention from EEG data in situations where music and speech are present. The results of this study indicate that linear regression can also be used for AAD when listening to music if the model is trained for musical signals.**

***Index Terms*— Auditory attention, electroencephalography (EEG), envelope tracking, stimulus reconstruction, music listening.**

I. Introduction

IN COMPLEX sound scenes, human beings have the ability to segregate sound streams and to focus their attention on one of the multiple sounds present [41]. A considerable amount of literature has been published on this ability, often called the cocktail party effect [10], [30]. These studies particularly focus on situations where multiple speech signals are presented to a listener. However, less research has explored this effect where music is also present.

The last two decades have seen a growing trend toward auditory attention decoding (AAD) from neuroimaging, as a way to understand the underlying mechanisms and as a potential application for future brain-computer interface (BCI) or neuro-steered hearing assisting devices [3], [6]. Auditory attention has been shown to induce neural responses [5], [13], [22], for example, by modulating some neural frequency bands [26], or by reshaping neural events [36], both during speech listening [27] or music listening [11], [25], [42]. These variations in event-related potentials (ERP) can be used to decode auditory attention [11], [28], [42], but they present several limitations. One of these limitations is the requirement of the specific onset of the auditory signal, while another challenge is the noisy nature of the neural signal. Those limitations raise the need for several repetitions of the task to extract useful ERP for hearing assistive BCI.

AAD has also been explored by looking at the mechanisms of neural entrainment. Some acoustic features of the audio heard, such as the temporal envelope of the audio signal heard by a listener, are tracked by the brain [20]. It led to new methods to analyze cortical responses due to continuous audio stimuli based on linear (or non-linear) models that estimate: either the neural response from the audio signal (encoding) [12]; or the audio signal from the cortical response (decoding) [2], [34], [35]. The decoding process, also known as the backward method or stimulus reconstruction method, has been demonstrated to be sensitive to auditory attention: when the listener focuses on one source of sound in a complex auditory environment, the cortical tracking of that attended sound is increased compared to the tracking of the unattended sounds [2], [12], [34], [44]. Several studies have demonstrated this influence of attentional factors, from magnetoencephalography [19], intracranial EEG [31] and intensively from scalp EEG [4], [6], [7], [21], [32], [33], [34]. Most of these studies are based on speech listening scenarios, where listeners have to solve cocktail party effects with often two competing speech streams.

The stimulus reconstruction approach has recently been successfully applied to reconstruct musical signals [15], [17],

Manuscript received 11 October 2022; revised 20 March 2023 and 13 June 2023; accepted 27 June 2023. Date of publication 30 June 2023; date of current version 12 July 2023. This work was supported in part by Bang & Olufsen A/S, Denmark; and in part by the Innovation Fund Denmark (IFD) under Grant 9065-00270B. *(Corresponding author: Adéle Simon.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Den Videnskabsetiske Komité for Region Nordjylland.

Adéle Simon and Søren Bech are with the Research Department, Bang & Olufsen A/S, 7600 Struer, Denmark, and also with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: amds@es.aau.dk; sbe@es.aau.dk).

Gérard Loquet is with the Department of Audiology and Speech Pathology, The University of Melbourne, Melbourne, VIC 3010, Australia (e-mail: gerard.loquet@unimelb.edu.au).

Jan Østergaard is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: jo@es.aau.dk).

Digital Object Identifier 10.1109/TNSRE.2023.3291239

[24], [29] but to a lesser extent than speech signals [24], [45]. AAD has also been applied to music, where the goal was to decode attention directed towards individual instruments in a multi-instrumental musical piece [3], [9], [23]. This new research focus is relevant as music is often present in natural sound scenes, either as a distracter or as a target of attention. Therefore, performance and potential specificities of the stimulus reconstruction method in a so-called "musical cocktail party scenario" could be explored in a context where multiple sounds that are speech or music are present and compete for the listeners' attention.

The present study investigated the performance of auditory attention decoding based on a linear stimulus reconstruction method in a musical cocktail party scenario. The primary goal of this study is to test the performance of an AAD based on previously used methodology in a listening situation that includes music. In the experiment, participants listened to a target sound, either speech or music, in the presence of a competing distracter sound, which was also either speech or music. During the listening task, the participants' cortical responses were continuously measured with high-density EEG and used to train a linear model that was then used to reconstruct the target stimuli and decode attention. The experimental strategy was designed to test the hypotheses that the temporal envelope of the target signal can be reconstructed regardless of whether it is speech or music, and with an accuracy above chance level; this decoding approach can be successfully used to decode attention in a musical cocktail party scenario.

II. Methods

A. Participants

Thirty-five participants (14 females), aged between 21 and 33-year-old (mean = 26,29) took part in the experiment. No participants reported a known history of neurological disorder or hearing loss. Apart from the three participants who were native English speakers, all the others had working experience or followed education in English. The participants were compensated for their participation, and written informed consent was obtained from all the participants. After recording, two participants were excluded from the data due to poor data quality in the raw data and thus not used for further analysis (due to the large number of artefacts that contaminate every trial).

B. Procedure

Each participant undertook 32 trials of one minute each. In each trial, they were presented with two different sound streams coming from separate loudspeakers. The loudspeakers were separated in space in front of the listener (+/− 30 ° azimuth). For each trial, participants were asked to pay attention to one of the sounds (the target), while ignoring the other sound (the distracter). The target, as well as the distracter, could be either speech or music (see figure 1).

Before starting the task, the participants did a training trial consisting of a trial similar to the real one with stimuli that were not later reused in the study. They had the opportunity to repeat this training as many times as they wanted and to ask questions about the task before starting. At the end of each trial, the participants answered two questions related to their attention level and the quality of their listening experience. Both questions were rated on a continuous scale with endpoint labels offset 1.5 cm after the start and before the end of a 15 cm long scale.

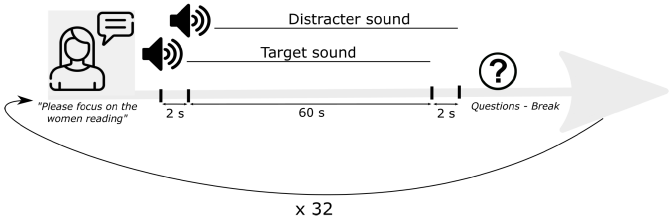

Fig. 1. A trial started with a visual cue that indicates which sound is the target. Right after, the two sound streams start with a 2s offset to help the participant to focus on the target. Participants listen to the two concurring sounds for 60s. At the end of each trial, participants have to answer two questions.

- *"How difficult was it to focus on the target stimuli?"* - Endpoint labels: *Easy* and *Difficult*
- *"How would you describe your listening experience?"* - Endpoint labels *Bad* and *Excellent*

The subjects could take a break between trials. The participants were instructed to keep their gaze fixed on a cross in the middle of the screen for the entire trial duration and asked to minimize body movements and blinking.

C. Stimuli

Four categories of stimuli were used, divided into two types (music and speech), with each type separated into two genres.

- Piano Music: 8 excerpts of mono instrumental pieces played on a piano
- Electronic music: 8 excerpts of polyphonic pieces of instrumental electronic music
- Speech female: 8 excerpts of an audiobook read by a woman in English
- Speech male: 8 excerpts of an audiobook read by a man in English

Each excerpt was one minute long, and the participants actively listened to the target throughout the whole minute. Participants listened to the same type of target for a full block (e.g. first a block of 8 trials of Piano Music, then a block of 8 trials of Speech female). The order of the block was randomized across participants. In the same trial, the target and the distracter could both be music, speech, or one of each type. Each excerpt was used only once as a target. Distracters were so that a balanced number of trials across conditions was obtained. For each trial, the distracter was randomly drawn from the pool of the relevant genre. In the case where both the target and distracter were music in a trial, the two excerpts could not belong to the same musical genre (e.g., target = piano music & distracter = piano music). The location of the target and the distracter (i.e. left or right loudspeaker) was randomized across trials.

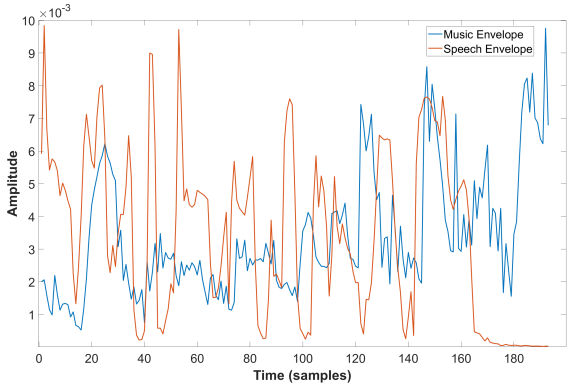

Fig. 2. Example of 3 seconds of an envelope for speech and music.

D. EEG Data Acquisition and Pre-Processing

The experiment was carried out in a single session for each participant. Continuous EEG data were recorded at 512 Hz using a 64-channel g.HIamp-Research system (g.tec Medical Engineering GmbH, Austria). The electrodes were placed on the scalp according to the 10-20 international system. The impedance of each electrode was maintained at lower than 5kOhms.

After data collection, pre-processing was carried out using EEGLAB v2021.1 [14]. The EEG data were referenced to the average of all scalp electrodes. The EEG channels contaminated by noise were visually inspected and interpolated from neighbouring electrodes. Independent Component Analysis (ICA) was run from EEGLAB and the automated detection plugin [37] allowed to remove the artefacts related to eye blinks or eye movements. The EEG data were bandpass filtered between 1 and 8 Hz and downsampled to a sampling rate of 64 Hz. The choice of cutoff frequency was based on previous studies on cortical stimulus reconstruction done on speech signal [34]. The influence of cutoff frequency was also tested on the present dataset: the results obtained support the importance of the 1-8 Hz frequency range for both speech reconstruction and music reconstruction. The trials where the artefacts were too significant were discarded (e.g., movements). The discarded data correspond to 7,68% of the total data.

For the signals, amplitude envelopes from both target and distracter were extracted using a Hilbert transform and then downsampled to the same sampling rate of 64 Hz. Examples of the envelopes for speech and music can be seen in Figure 2. The shape of envelopes for speech and music differ due to the nature of the signal: for speech signal, due to the pause between words, the envelope tends to drop to zero and show a greater depth of modulation compared to music envelopes.

E. Stimulus Reconstruction and Attention Decoding

The decoding of auditory attention from the EEG signal was done with a conventional stimulus reconstruction method [2], [12], [34]. With this method, the EEG signal is used to reconstruct an estimate of the input stimulus through a linear reconstruction multi-delay model. This model maps

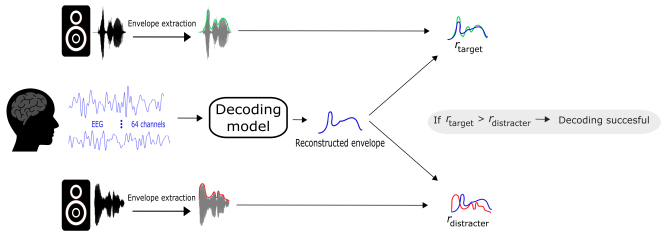

Fig. 3. Schematic summary of auditory attention decoder.

the cortical activity measured with the EEG to the stimulus envelope as follows:

$$s'(t) = \sum_{n} \sum_{\tau} g(\tau, n) R(t - \tau, n) \tag{1}$$

where s' is the reconstructed envelope, $R(t-\tau, n)$ is the EEG response at time $(t-\tau)$ for electrode n, and g is the linear model, which is a function of electrode n and time lags τ. The time lags τ cover the interval from 0 ms to 500 ms post-stimulus, in order to take into account time lags that have been shown to influence AAD for both speech [34] and Music [24]. The model g can be estimated by minimizing the mean squared error between the original and the reconstructed envelopes, which can be solved analytically using ridge regularization methods [44]:

$$g = (R^T R + I\lambda)^{-1} R^T S \tag{2}$$

where I is the identity matrix, and λ is the regularization parameter used to prevent overfitting [2], [44]. The hyperparameter λ was estimated through a cross-validation approach, as described in [12]. This test was run for each separated subset of data (target = speech female, target = speech male, target = music electronic, and target = music piano), in order to ensure that the regularization factor is optimized for each stimulus type. For those four categories, the optimal regularization factor that produced the highest reconstruction accuracy was similar at $\lambda = 10^5$.

The reconstruction accuracy is measured by calculating Pearson's r, the correlation coefficient, between the original target envelope and the reconstructed one (r_{target}). The correlation between the reconstructed envelope and the envelope extracted from the distracter was also calculated ($r_{distracter}$). The correlation is calculated with an entire trial, corresponding to 60 seconds of the reconstructed envelope and 60 seconds of the original envelope.

For each reconstruction, the attention decoding was evaluated by comparing the two correlation coefficients. A trial was considered successfully decoded if the reconstructed envelope had a greater correlation with the target envelope compared to the correlation with the distracter envelope (i.e., $r_{target} > r_{distracter}$).

For the present study, the stimulus reconstruction approach was made using a custom-made analysis script on Matlab R2021a.

F. Model Training

For each trial, a leave-one-out cross-validation method was used to train the models. Each trial was decoded using a model

obtained by averaging the parameter of the models trained on all other trials. Through this experiment, several types of models were used, all created with different sets of training data:

- Trained on all: All trials, both when the target is speech and music, minus the one under test, are used to calculate the models.
- Trained on the same type and the same genre: Congruent model, where all trials where the target is of the same type and the same genre as the one under test (e.g., Piano music), minus the one under test, are used to calculate the models.
- Trained on the same type: All trials where the target is of the same type as the one under test (either music or speech), minus the one under test, are used to calculate the models.
- Trained on opposite type: All trials where the target is not of the same type as the one under test are used to calculate the models.

III. Results

Two measures were used to assess the performance of auditory attention detection. The first one is the success rate of attention detection, which corresponds to the percentage of trials that were successfully decoded. To that aim, the correlation between the reconstructed envelope is compared to either the target's envelope (r_{target}) or the distracter's envelope ($r_{distracter}$). A trial is successfully decoded when the correlation of the reconstructed envelope with the target is greater than with the distracter ($r_{target} > r_{distracter}$). This success rate can indicate if the model allows decoding auditory attention better than chance. Chance level is calculated by taking the mean and the confidence interval of a binomial distribution with a success chance of 50%, corresponding to a random binary decision.

Following that, the reconstruction accuracy was also investigated, which corresponds to Pearson's correlation coefficients between the reconstructed envelope and the target envelope. The goal is to investigate if the linear model can reconstruct the target envelope better than chance, and then explore potential differences between the reconstruction of musical envelopes compared to the reconstruction of speech envelopes.

To establish the chance level, all conditions were compared with a "random reconstruction accuracy". The random reconstruction accuracy was calculated with a reconstructed envelope and an unrelated original envelope: e.g. the envelope of the target used for trial 1 of Subject 1, where the target was piano music, was correlated with the envelope reconstructed from trial 14 from Subject 6, where the target was female speech. The pairing between the original and reconstructed envelopes was randomized. Following that, permutation tests were used to compare the reconstruction accuracy for each condition to the random reconstruction accuracy with 10 000 permutations. For each condition, sample sizes used for the calculation of actual accuracy and random accuracy were equal.

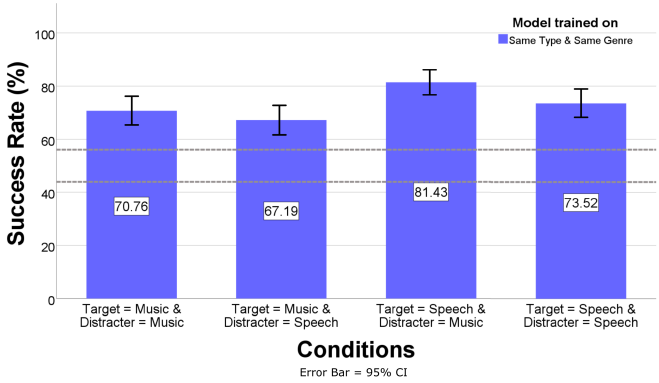

Fig. 4. Success rates across conditions obtained with models trained on the same type and same genre as the target - Chance level is indicated by dashed grey lines.

A. Congruent Model

The models were first calculated with training data picked only from trials where the target was similar to the trial under test, i.e. of the same type and same genre. By using condition-specific reconstruction filters, the assumption is that the models were not influenced by other listening conditions and would be fitted to each specific trial. The highest decoding success rate is obtained when the target of attention is speech, either when the distracter is music (Success rate = 81.43%) or when the distracter is speech (Success rate = 73.52%). When the listener is actively listening to music, the success rate of the AAD is a bit lower, at 70.76% when the distracter is also music or 67.19% when the distracter is speech.

1) Success Rate: As shown in Figure 4, when using a congruently trained model, auditory attention can be successfully decoded, above chance level (chance level interval = 43.89% to 56.11%), for all listening conditions.

2) Reconstruction Accuracy: For all conditions, reconstruction accuracy was significantly higher than the random reconstruction accuracy ($p < 0.0001$), suggesting that for all listening conditions, stimulus reconstructions are feasible above chance level.

To explore differences between reconstruction accuracy across conditions, Anova based on a general linear model was performed to explore the main effect of fixed (listening conditions). Participants were included as a random factor.

When comparing across conditions, the ANOVA shows a significant effect of listening conditions ($F(3, 998) = 25.980$, $p < 0.001$), with significant differences, calculated by posthoc comparisons with Bonferroni corrections, between both of the "target music" conditions and the "target speech" conditions ($p < 0.001$). Significant differences were also found between the two "target speech" conditions ($p = 0.004$). These results suggest that the stimulus reconstruction approach can better reconstruct the target stimulus when the listener is listening to speech compared to situations where the listener is listening to music, especially when the distracter is a musical sound (see figure 6-A).

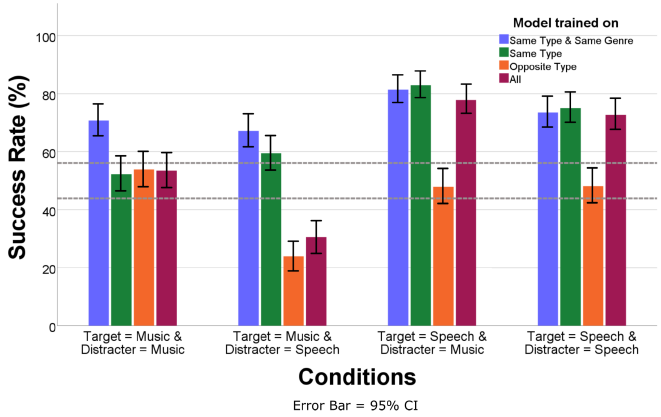

Fig. 5. Success rates across conditions obtained with differently trained models - Chance level is indicated by dashed grey lines.

B. Other Training Conditions

In a second analysis, the training of the model was also considered to evaluate the success rate and reconstruction accuracy when using models trained on various types of sounds. The goal of this analysis is to explore the influence of the training data on the performance of the model.

1) Success Rate: Figure 5 shows the success rate of auditory attention decoding when using models that are differently trained. For the "target speech" conditions, the success rate appears to be unaffected by the training data used, apart from the "trained on opposite type" condition. In that condition, the auditory attention model performs around the chance level, suggesting that in such a case, auditory attention cannot be successfully detected. For the other conditions, when the target of attention is music, the training of the model influences the performances. For the "target music & distracter music" condition, only the most congruent condition (trained on the same type and same genre) allows for successful decoding, while all the other training conditions perform around the chance level. For the "target music & distracter speech" condition, while both congruent training conditions perform above chance level (trained on the same type and same genre and trained on the same type), for the two other conditions, the success rate is considerably low. The low success rate, below chance level, suggests that in these cases, the model tends to reconstruct the distracter better than the target. The poor success rate observed in this condition suggests that a general decoder, trained on both speech and music, tends to be biased toward speech reconstruction.

2) Reconstruction Accuracy: As for the congruent models, reconstruction accuracy was significantly higher than the random reconstruction accuracy ($p < 0.0001$) for all conditions and for both target reconstruction accuracy and distracter reconstruction accuracy. It suggests that for all listening conditions, stimulus reconstructions are feasible above chance level, either with the target or with the distracter.

A three-way ANOVA based on general linear model was calculated with both listening conditions and model training as fixed factors and participants as a random factor. Main effects and interactions between fixed factors (Listening Condition x Model Training) were explored, and residuals were checked to be normally distributed. Results showed significant effects of both Listening conditions ($F(3, 4088) = 118.293$, $p < 0.001$) and Training conditions ($F(3, 4088) = 96.642$, $p < 0.001$). A significant interaction between the two factors was also found ($F(9, 4088) = 7.065$, $p < 0.001$). (See figure 6-A for results) These results suggest that when the target of attention is speech, the reconstruction is less precise when using a model trained on different types of signals (here trained only on music). For all other training conditions, reconstruction performances are equivalent. When music is the target of attention, the training of the model influences the results. Training on an incongruent model, with only trials from a different type, decrease the reconstruction performance for speech. Contrary to the "target speech" conditions, here, using a model trained on the same musical genre increase the reconstruction performance compared to a generic musical model ("trained on the same type") or a generic model ("trained on all").

Figure 6-B shows the correlations obtained when calculating the correlation between the reconstructed envelope and the distracter envelope. First, it can be observed that the results obtained are above chance level, for all conditions. That suggests that it is also possible to reconstruct sounds that were heard but not actively focused on. However, for most situations, this reconstruction is lower, as the correlation with the original envelope is smaller, indicating that the distracter sound can also be reconstructed, but to a lesser extent than the target.

The situation is different when the target is music and the distracter is speech. In that case, when the model is trained only on musical signal ("trained on the same type" or trained on the same type and same genre"), the reconstruction accuracy for the distracter is lower than the reconstruction accuracy with the target, and follow a similar trend compared to the other listening conditions. When the model is trained on speech signal ("trained on opposite signal" or "trained on all"), the reconstruction accuracy of the distracter is greater. It is still lower than the target reconstruction accuracies obtained with speech as a target, but on average greater than the target reconstruction accuracy obtained with music as a target. This suggests that when trained with speech signals, the model may be biased towards speech reconstruction, which can also explain the poor success rate obtained for the conditions where music is a target, speech is a distracter, and the model is trained on speech.

C. Effect of Size of Training Set

In the aforementioned analysis, the size of the training set differed: as the general model used all available data, the training dataset is larger than for the model trained on only a subset of data (e.g. congruent models trained on one specific type and genre). This approach was chosen to optimize training by using as much data as possible. However, this difference in the size of the training sets may influence the conclusions of the current study. To control this factor, the success rates for AAD were recalculated with models trained on smaller training sets to ensure that all training subsets were of equal size (i.e. the models were trained on 252 trials of one

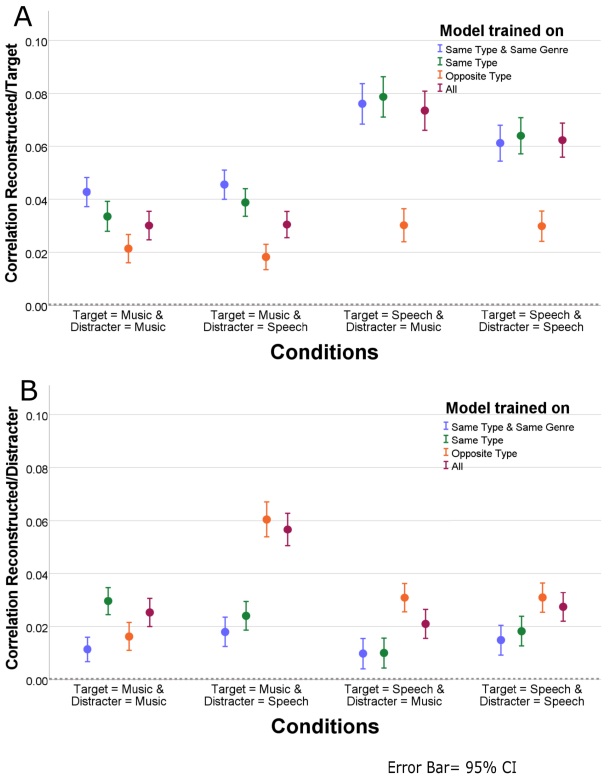

Fig. 6. Reconstruction accuracies across conditions obtained with differently trained models, between A- reconstructed and target stimuli; B- reconstructed and distracter stimuli - Chance level is indicated by dashed grey lines.

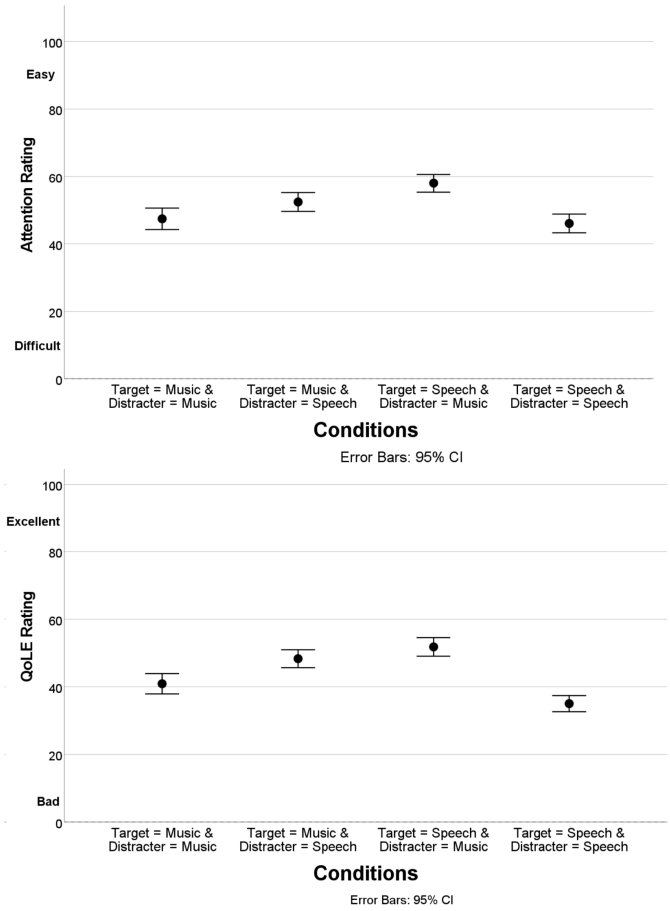

Fig. 7. Subjective ratings across conditions, mean and 95% CI across participants.

minute). For each training condition, the training subsets were randomly selected from the available training data. During the selection, the type of data was controlled to ensure a balanced distribution of conditions in the smaller training set (e.g. for the subset selection for a model trained on both speech and music, the number of trials where Target = Music is equal to the number of trials where Target = speech). An exact McNemar's test was used to test if the success rates obtained with equally-sized training sets differ from the success rates from unequally-sized training sets. The test determined that there were no significant differences between the two conditions (p =.275).

D. Subjective Ratings

Two-way mixed model ANOVAs were conducted to examine the effect of condition and participants (included as a random factor), as well as the interactions, on the subjective ratings of the participants, attention and quality of listening experience (QoLE); residuals were checked to be normally distributed.

For attention, significant effects were found for the condition factor $F(3, 998) = 20.081, p < 0.001$. For QoLE, significant effects were found for the condition factor $F(3, 998) = 48.618, p < 0.001$. This suggests that while there are differences between the conditions in terms of difficulty to focus on the target stimuli and quality of listening experience, it also varies across individuals. Results can be seen in figure 7.

In order to explore a potential link between cortical reconstruction accuracy and subjective ratings, Pearson's correlations have been run between the reconstruction accuracy values and the subjective rating. For the attention ratings, a small but significant correlation has been found with the reconstruction accuracy ($r(5168) = 0.052, p < 0.001$). Similarly, the ratings of QoLE ratings are slightly correlated with the cortical reconstruction accuracy ($r(5168) = 0.042$, $p = 0.003$)

IV. Discussion

This study attempts to decode auditory attention from continuous cortical responses, measured with EEG, in a musical cocktail party scenario. Participants were presented with two streams of sounds simultaneously, which could be either speech or music, and asked to focus on one of the sounds (target) while ignoring the other (distracter). A linear regression method that maps the cortical data to the audio signal was used to reconstruct the input stimuli, and the reconstruction was used to decode the attention of the listener.

For each trial, the attention decoding was done by comparing the reconstructed envelope with both the target envelope and the distracter envelope. As the stimulus reconstruction approach has been shown to be sensitive to selective auditory attention, it was hypothesized that the reconstructed envelope should correlate better with the envelope of the target stimulus

compared to the correlation with the distracter stimulus, irrespective of the type of target of distracter (music or speech).

When tested on a congruently trained model (i.e., trained and tested on trials with the target of the same type and genre), auditory attention can be successfully decoded for all listening conditions above chance level. The type of data used for the training of the model also impacted the success rate of the decoding. When the target of attention was speech, all training conditions led to successful decoding, with an equivalent success rate, except for one condition. If the model was trained only on musical trials and tested on speech, the success rate dropped around the chance level. In this condition, the decoding was unsuccessful.

When the target of attention was music, results also vary across conditions depending on the type of distracter. If the target was music, and the distracter was also music, only the congruent training (same type & same genre), leads to successful decoding. All the other training conditions perform around chance level. However, when the target was music and the distracter was speech, both congruently trained models (model trained on the same type & same genre, as well as model same type) perform above chance level. However, when the training set also included speech data, the success rate dropped considerably. For the "trained on all" and "trained on opposite type" (i.e., here trained on speech), the success rate was below chance level (23.95 - 30.53%), which suggests that the model reconstructs the distracter better than the target. This finding was unexpected and suggests that, in these conditions, the model is more influenced by aspects specific to speech than by aspects related to auditory attention.

For all listening conditions, the values of reconstruction accuracies obtained are better than chance. It suggests that, even when multiple sounds are present in a sound scene, the linear regression approach can be used to reconstruct the stimulus that a listener is attentive to, both when this stimulus is speech and music. These results, especially the order of magnitude of the reconstruction accuracy obtained through this study, are consistent with previous work on auditory attention decoding using linear regression during a cocktail party scenario with only speech [21], [32], [34], or previous work using this reconstruction approach during music listening [9], [23], [29]. Reconstruction accuracies are overall greater when the target of attention is speech compared to the trials where the target of attention is music. This difference in tracking accuracies matched those observed by [45]. In addition, the results showed that the training of the model can influence the reconstruction accuracy. An incongruently trained model (e.g., a model trained on music and tested on speech), significantly reduces the reconstruction accuracies. When the target of attention is music, the reconstruction accuracy seems to be more sensitive to the training of the model: The reconstruction accuracy is significantly higher when decoded with a model trained only on the same musical genre, compared to the other models (i.e., trained only on music trials, but also including other musical genres, or trained on both speech and music). Nevertheless, it is coherent with the results obtained for the success rate of attention decoding. This difference might be greater than the difference due to the auditory attention, leading to a better reconstruction of speech than music, even if the attention was directed to speech.

The difficulty of the task and the difficulty of attending to the target stimulus might influence the decoding performance, as the attention of the listener might not be perfectly on the target during a challenging trial. However, the correlation between subjective ratings of attention and reconstruction accuracy has been found to be small. This difficulty in attending to the target might explain the small decrease in performance for speech listening in presence of speech compared to speech listening in presence of music. It is, however, different for the situation where the target of attention is music and the distracter speech. While being rated easier than the speech on speech or music on music situation, the decoding performances are worse. More research would be needed to explore the relationship between cortical reconstruction accuracy, the difficulty of the task and listening effort. The fact that participants were not native English speakers may also influence the neural response (and thus the reconstruction accuracy), or modulate how they attend to the speech signal [38].

The choice of the bandpass filter (1 to 8 Hz) might also have influenced the reconstruction accuracy, as it has been shown that the cortical tracking of sound differs between speech and music listening at lower frequencies [45]. This study aimed to test an AAD method, that has previously proven successful on speech listening, on a music listening task. However, that means that the method and the parameters have been optimized for speech and may be sub-optimal for music. Further research should be undertaken to investigate how the performances are influenced by some of the signal or model parameters, such as filtering of the EEG data and choice of audio features used in the AAD. In order to increase reconstruction and attention decoding performances for music listening, further work could also explore stimulus reconstruction based on other audio features than the envelope, which may be better tracked by the brain during music listening such as mel spectrogram [9] or notes onset timing [29]. For music listening, it also has been suggested that spectral modulation plays a greater role than temporal modulations [43]. The open questions that arose after this study are: to what extent do the parameters used for the models (such as the audio features) maximize the reconstruction for speech compared to music; and if there are other parameters that can be more suited for music reconstruction?

The differences between speech and music conditions, both in reconstruction accuracy and success rate, could also be due to separate cortical processes for speech or music listening involving different parts of the brain [1]. In the context of stimulus reconstruction, when trying to maximize reconstruction accuracy, differences were found between speech and music listening for the optimal latencies [40] and the selection of the electrode [39]. These results suggest temporal and topographic variations that could indicate the presence of differences in cortical processes activated during speech or music listening. Additional studies would be needed to investigate further the temporal and topographic variations during cortical music and speech processing.

Furthermore, the reconstruction accuracy results might be due to an enhanced cortical tracking for speech compared to music, which would be in line with the recent findings by [45]. Differences could be related to some brain processing specific to speech, or they could be due to some acoustical aspects specific to speech signals. Another potential explanation, as suggested by [45], is that these differences may be linked to some high-level speech-specific features, such as phoneme processing [16] or phonotactic probabilities [18] or semantics aspects [8] that could increase the cortical tracking of speech and thus lead to greater reconstruction accuracy. Aspects related to the signal itself might influence the differences obtained for signal reconstruction. The shape of the envelopes differs between speech and music. For speech, due to the pause between words, the envelopes go down to zero, followed by sharp jumps to high amplitude a the beginning of a new word. For the music envelope, the envelopes are more "flat", resulting in a lower dynamics range (see Figure 2). Further studies, which take these variables into account, would be needed to get a better understanding of cortical auditory tracking, and variations between speech and music, in order to untangle how different aspects of speech influence the reconstruction accuracy.

Independent of the underlying factors influencing the reconstruction accuracy, the present results should be considered when designing a versatile AAD, especially in cases where both speech and music could be the target of attention. Due to the difference in reconstruction accuracy observed between speech and music reconstruction, the current decision criteria (i.e., comparing the reconstruction accuracy for both target and distracter) may not be suitable as it does not take into account these differences. Other approaches, for instance, using thresholds to classify between target or distracter, might be more appropriate (e.g., the reconstruction accuracy should be above a music threshold to be considered as a target of attention if there is music in the sound scene). Thresholds could also be used to correct the bias toward speech found in the general model (i.e., trained on both speech and music) to ensure that the apriori probability for a trial to be classified as speech or music is equal. While this approach would be interesting to develop AAD, it is outside the scope of the present study to determine such thresholds: more data and more diverse situations would be needed to explore and determine relevant thresholds. In addition, for AAD implementation, using threshold would require knowledge about the sound scene when using the AAD, such as an acoustic scene classification to inform about the presence of speech or music in the sound scene to decode.

Overall, this study shows that auditory attention decoding is feasible for musical cocktail party scenarios, both during active speech listening and active music listening. However, for music listening, the decoding model needs to be fitted to the target stimulus. This is a limitation for the potential applications of such technologies, as it would be necessary to know the type of stimuli present in the sound scene to apply the right model. Future work could further explore the differences between speech listening and music listening for auditory attention decoding to gain knowledge about the underlying mechanisms for both music listening and speech listening and attempt to improve the performance in AAD during music listening.

REFERENCES

[1] P. Albouy, L. Benjamin, B. Morillon, and R. J. Zatorre, "Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody," *Science*, vol. 367, no. 6481, pp. 1043–1047, Feb. 2020.

[2] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers Neurosci.*, vol. 13, p. 153, Mar. 2019.

[3] W. W. An et al., "Decoding music attention from 'EEG headphones': A user-friendly auditory brain-computer interface," in *Proc. ICASSP*, 2021, pp. 985–989.

[4] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of different acoustic components on EEG-based auditory attention decoding in noisy and reverberant conditions," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 652–663, Apr. 2019.

[5] C. L. Baldwin, *Auditory Cognition and Human Performance: Research and Applications*. Boca Raton, FL, USA: CRC Press, 2012.

[6] J. Belo, M. Clerc, and D. Schön, "EEG-based auditory attention detection and its possible future applications for passive BCI," *Frontiers Comput. Sci.*, vol. 3, Apr. 2021, Art. no. 661178.

[7] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.

[8] M. P. Broderick, N. J. Zuk, A. J. Anderson, and E. C. Lalor, "More than words: Neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative," *Eur. J. Neurosci.*, vol. 56, no. 8, pp. 5201–5214, 2022.

[9] G. Cantisani, S. Essid, and G. Richard, "EEG-based decoding of auditory attention to a target instrument in polyphonic music," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 80–84.

[10] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[11] I. Choi, S. Rajaram, L. A. Varghese, and B. G. Shinn-Cunningham, "Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography," *Frontiers Hum. Neurosci.*, vol. 7, p. 115, Apr. 2013.

[12] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers Hum. Neurosci.*, vol. 10, p. 604, Nov. 2016.

[13] S. Crottaz-Herbette and V. Menon, "Where and when the anterior cingulate cortex modulates attentional response: Combined fMRI and ERP evidence," *J. Cognit. Neurosci.*, vol. 18, no. 5, pp. 766–780, May 2006.

[14] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[15] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate decoding of imagined and heard melodies," *Frontiers Neurosci.*, vol. 15, Aug. 2021, Art. no. 673401.

[16] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biol.*, vol. 25, no. 19, pp. 2457–2465, Oct. 2015.

[17] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. De Cheveigné, "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," *Acoust. Sci. Technol.*, vol. 41, no. 1, pp. 361–364, 2020.

[18] G. M. Di Liberto, D. Wong, G. A. Melnik, and A. De Cheveigné, "Low-frequency cortical responses to natural speech reflect probabilistic phonotactics," *NeuroImage*, vol. 196, pp. 237–247, Aug. 2019.

[19] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 29, pp. 11854–11859, Jul. 2012.

[20] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: Functional roles and interpretations," *Frontiers Human Neurosci.*, vol. 8, p. 311, May 2014.

[21] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, Aug. 2017.

[22] J. C. Hansen and S. A. Hillyard, "Endogeneous brain potentials associated with selective auditory attention," *Electroencephalogr. Clin. Neurophysiol.*, vol. 49, nos. 3–4, pp. 277–290, Aug. 1980.

[23] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating cortical instrument representations during auditory stream segregation and integration with polyphonic music," *Frontiers Neurosci.*, vol. 15, Sep. 2021, Art. no. 635937.

[24] L. Hausfeld, L. Riecke, G. Valente, and E. Formisano, "Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes," *NeuroImage*, vol. 181, pp. 617–626, Nov. 2018.

[25] N. J. Hill and B. Schölkopf, "An online brain-computer interface based on shifting attention to concurrent streams of auditory stimuli," *J. Neural Eng.*, vol. 9, no. 2, Apr. 2012, Art. no. 026011.

[26] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a 'cocktail party,"' *J. Neurosci.*, vol. 30, no. 2, pp. 620–628, Jan. 2010.

[27] D. J. Lee, H. Jung, and P. Loui, "Attention modulates electrophysiological responses to simultaneous music and language syntax processing," *Brain Sci.*, vol. 9, no. 11, p. 305, Nov. 2019.

[28] P. Loui, T. Grent, D. Torpey, and M. Woldorff, "Effects of attention on the neural processing of harmonic syntax in western music," *Cognit. Brain Res.*, vol. 25, no. 3, pp. 678–687, Dec. 2005.

[29] G. Marion, G. M. D. Liberto, and S. A. Shamma, "The music of silence: Part I: Responses to musical imagery encode melodic expectations and acoustics," *J. Neurosci.*, vol. 41, no. 35, pp. 7435–7448, 2021.

[30] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, Sep. 2005.

[31] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[32] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, "Target speaker detection with concealed EEG around the ear," *Frontiers Neurosci.*, vol. 10, p. 349, Jul. 2016.

[33] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications," *J. Neural Eng.*, vol. 12, no. 4, Aug. 2015, Art. no. 046007.

[34] J. A. O'Sullivan et al., "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015.

[35] B. N. Pasley et al., "Reconstructing speech from human auditory cortex," *PLoS Biol.*, vol. 10, no. 1, Jan. 2012, Art. no. e1001251.

[36] T. W. Picton and S. A. Hillyard, "Human auditory evoked potentials. II: Effects of attention," *Electroencephalogr. Clin. Neurophysiol.*, vol. 36, pp. 191–200, Jan. 1974.

[37] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, Sep. 2019.

[38] R. Reetzke, G. N. Gnanateja, and B. Chandrasekaran, "Neural tracking of the speech envelope is differentially modulated by attention and language experience," *Brain Lang.*, vol. 213, Feb. 2021, Art. no. 104891.

[39] A. Simon, S. Bech, G. Loquet, and J. Østergaard, "Electrodes selection for cortical auditory attention decoding with EEG during speech and music listening," in *Proc. 25th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2022, pp. 1–6.

[40] A. M. D. Simon et al., "Optimal time lags for linear cortical auditory attention detection: Differences between speech and music listening," in *Proc. 19th Int. Symp. Hearing, Psychoacoust., Physiol. Hearing, Auditory Modelling, Ear Brain*. Zenodo, 2022.

[41] E. S. Sussman, "Auditory scene analysis: An attention perspective," *J. Speech, Lang., Hearing Res.*, vol. 60, no. 10, pp. 2989–3000, Oct. 2017.

[42] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification," *J. Neural Eng.*, vol. 11, no. 2, Apr. 2014, Art. no. 026009.

[43] I. Wollman, P. Arias, J.-J. Aucouturier, and B. Morillon, "Neural entrainment to music is sensitive to melodic spectral complexity," *J. Neurophysiol.*, vol. 123, no. 3, pp. 1063–1071, Mar. 2020.

[44] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers Neurosci.*, vol. 12, p. 531, Aug. 2018.

[45] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLOS Comput. Biol.*, vol. 17, no. 9, Sep. 2021, Art. no. e1009358.