No Evidence of Attentional Modulation of the Neural Response to the Temporal Fine Structure of Continuous Musical Pieces

Octave Etard[1], Rémy Ben Messaoud[1], Gabriel Gaugain[1], and Tobias Reichenbach[2]

Abstract

■ Speech and music are spectrotemporally complex acoustic signals that are highly relevant for humans. Both contain a temporal fine structure that is encoded in the neural responses of subcortical and cortical processing centers. The subcortical response to the temporal fine structure of speech has recently been shown to be modulated by selective attention to one of two competing voices. Music similarly often consists of several simultaneous melodic lines, and a listener can selectively attend to a particular one at a time. However, the neural mechanisms that enable such selective attention remain largely enigmatic, not least since most investigations to date have focused on short and simplified musical stimuli. Here, we studied the neural encoding of classical musical pieces in human volunteers, using scalp EEG recordings. We presented volunteers with continuous musical pieces composed of one or two instruments. In the latter case, the participants were asked to selectively attend to one of the two competing instruments and to perform a vibrato identification task. We used linear encoding and decoding models to relate the recorded EEG activity to the stimulus waveform. We show that we can measure neural responses to the temporal fine structure of melodic lines played by one single instrument, at the population level as well as for most individual participants. The neural response peaks at a latency of 7.6 msec and is not measurable past 15 msec. When analyzing the neural responses to the temporal fine structure elicited by competing instruments, we found no evidence of attentional modulation. We observed, however, that low-frequency neural activity exhibited a modulation consistent with the behavioral task at latencies from 100 to 160 msec, in a similar manner to the attentional modulation observed in continuous speech (N100). Our results show that, much like speech, the temporal fine structure of music is tracked by neural activity. In contrast to speech, however, this response appears unaffected by selective attention in the context of our experiment. ■

INTRODUCTION

Music is a fascinatingly complex acoustic stimulus. Listeners can follow the multiple melodic lines played by different instruments composing an orchestra or listening to the ensemble as a whole. This demonstrates one of the fundamental properties of the auditory system—its ability to segregate concurrent and sequential sounds and group them into perceptual streams (Bregman, 1994). The formation of these percepts is influenced by the physical characteristics of the incoming sounds, including notably pitch, location, onset synchrony, and presentation rate (Moore & Gockel, 2012; Cross, Hallam, & Thaut, 2008), but also by the listener's exposure to the stimulus regularities and learnt patterns (Winkler, Takegata, & Sussman, 2005; Winkler, Teder-Sälejärvi, Horváth, Näätänen, & Sussman, 2003). These intricate characteristics are thought to rely on a distributed network from the periphery to the auditory cortex (Kashino & Kondo, 2012), with invasive animal studies suggesting that neurons as early as the cochlear nucleus exhibit response curves consistent with human behavioral (Pressnitzer, Sayles, Micheyl, & Winter, 2008; Pressnitzer, Meddis, Delahaye, & Winter, 2001). However, the relative contributions of the different auditory structures to the formation of auditory percepts, as well as where in the auditory pathways these emerge, remain unclear. This is partly due to the difficulties that arise when comparing human psychophysics studies with invasive animal neural recordings. Human noninvasive neuroimaging studies provide a valuable tool to bridge that gap. However, many of these studies have focused on well-controlled but simplified patterns that do not encompass the much richer structure of real-world acoustic signals. Combining complex continuous stimuli, neuroimaging, and statistical models may further our understanding of the neural mechanisms underlying auditory scene analysis, in particular when higher cognitive constructs, such as selective attention, are involved.

Recent studies have indeed shown how these methods can relate key features of speech to electrophysiological recordings and inform on the neural mechanisms of speech processing (Wöstmann, Fiedler, & Obleser, 2017; Di Liberto, O'Sullivan, & Lalor, 2015; Ding & Simon, 2012a, 2014). For example, cortical activity has been found

[1]Imperial College London, [2]Friedrich-Alexander University Erlangen-Nuremberg

https://doi.org/10.1162/jocn_a_01811

to track slow (<8 Hz) amplitude fluctuations in speech (Ding & Simon, 2012b; Pasley et al., 2012; Lalor & Foxe, 2010; Nourski et al., 2009). This envelope tracking emerged in the primary auditory cortex and was found to be modulated by selective attention in competing speakers and speech-in-noise experiments, thereby demonstrating a neural correlate of independent source representation in the cortex (O'Sullivan et al., 2015; Ding & Simon, 2012a). In a similar manner, subcortical as well as, presumably to a lesser degree, cortical responses were shown to emerge to the higher-frequency (>80 Hz) stimulus structure (Etard, Kegler, Braiman, Forte, & Reichenbach, 2019; Bidelman, 2018; Maddox & Lee, 2018; Forte, Etard, & Reichenbach, 2017; Coffey, Herholz, Chepesiuk, Baillet, & Zatorre, 2016). Moreover, a recent study showed that subcortical neural responses to the pitch of continuous speech are stronger when the stimulus is attended rather than ignored (Forte et al., 2017). This result suggests that the pitch of a speaker could be used by the brain to perceptually segregate the speech signal from background noise, a finding that agrees with previous psychophysical studies that have found it easier to differentiate two concurrent speech signals if their fundamental frequencies differ (Madsen, Whiteford, & Oxenham, 2017; de Cheveigné, Kawahara, Tsuzaki, & Aikawa, 1997).

Understanding how the brain can focus on a single instrument among others relates to speech through a major challenge in auditory neuroscience, the cocktail party problem. This problem acquired its name from the observation that humans do remarkably well at understanding a target speaker in a noisy environment, such as in a busy restaurant or in a loud bar (Haykin & Chen, 2005; Cherry, 1953). The temporal fine structure of speech originates from the periodic opening and closing of the vocal folds at the so-called fundamental frequency. The spectrum of these voiced speech parts is therefore dominated by the fundamental frequency as well as its many higher harmonics, leading to a pitch perception in the listeners. Musical tones are similarly characterized by a fundamental frequency and higher harmonics, resulting in a characteristic temporal structure that causes a pitch perception. The proximity of fundamental frequencies of subsequent tones has been found to aid the formation of an auditory stream (Oxenham, 2008; Bregman, Liao, & Levitan, 1990). Consequently, just as the neural tracking of temporal fine structure could help listeners attend to a voice in background noise, such a neural mechanism may aid with attending to a particular melodic line (Micheyl & Oxenham, 2010).

Here, we investigated this hypothesis by using linear models to assess neural responses to the temporal fine structure of continuous melodic lines. To establish that we could measure neural responses to continuous musical pieces, we first presented volunteers with continuous classical Bach pieces consisting of melodic line played by one instrument while recording their brain activity through a bipolar EEG montage. We related the neural activity to the stimulus waveforms using encoding and decoding methods. To then assess a putative effect of selective attention on these neural responses, we also presented the volunteers with continuous pieces consisting of two different melodic lines played by competing instruments, a guitar and a piano. Participants were asked to selectively attend to one of the two lines, and we contrasted the neural responses to each instrument when it was attended to when it was ignored.

Because of multiple nonlinearities in the auditory periphery, both the temporal fine structure and the envelope of the stimuli are represented in neural responses. This encoding has traditionally been investigated in humans by studying time-locked responses to transient or periodic features of repeated short sound tokens such as clicks, pure or complex tones, syllables, and words (Skoe & Kraus, 2010). These paradigms typically present a particular stimulus as well as its opposite waveform many times. The neural responses to each polarity are then summed to emphasize responses to the envelope or subtracted to emphasize response to the temporal fine structure (Krizman & Kraus, 2019; Aiken & Picton, 2008). Here, we used continuous, long stimuli to derive auditory neural responses to their temporal fine structure using linear convolutive models.

METHODS

Experimental Design and Statistical Analysis

Seven of Bach's "Two-Part Inventions" (BWV 772-801) were used in this study. Each Two-Part Invention is a short keyboard composition that consists of two melodic lines: one played by the left hand, and the other played by the right. We synthesized the stimuli in GarageBand (Apple) from Musical Instrument Digital Interface (MIDI) files, with the left hand being played by a piano and the right hand being played by a guitar. To assess the attention of participants to a particular melodic line, vibratos were inserted in both lines.

Volunteers were presented with two types of stimuli. The first type, "Single Instrument" (SI), consisted of one single instrument, piano or guitar, that played one melodic line. The second type, "Competing Instruments" (CI), contained both melodic lines of a Two-Part Invention, one played by the piano and the other played by the guitar.

The different stimuli were presented in blocks (Figure 1). Each block contained one SI stimulus and one subsequent CI stimulus, both of which were obtained from the same Two-Part Invention. During the CI stimulus, the volunteers were asked to selectively listen to the instrument that they heard before in the SI stimulus. They were also asked to identify the vibratos embedded into that melodic line.

Blocks with the SI stimulus played by the piano alternated with those played by the guitar. Each of the seven

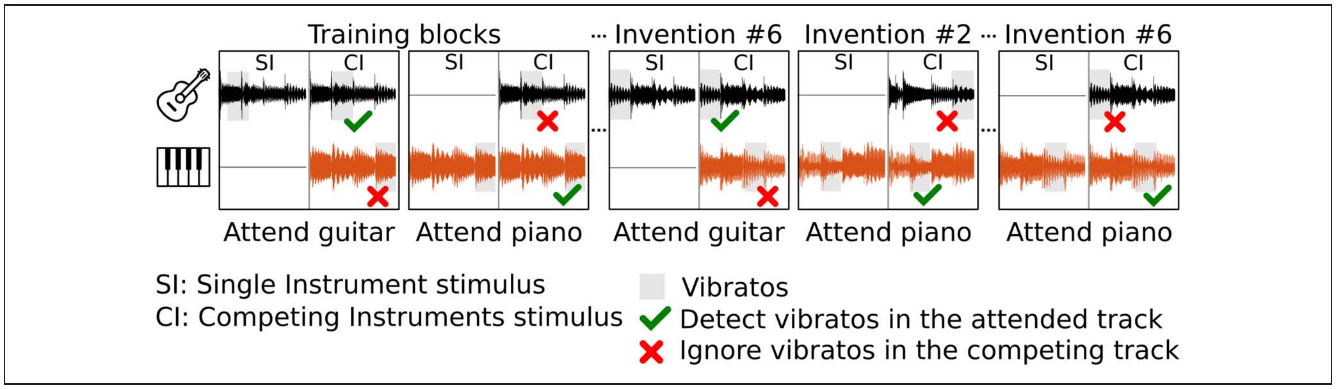
Training blocks
Invention #6
Invention #2
Invention #6
SI
CI
Attend guitar
Attend piano
Attend guitar
Attend piano
Attend piano
SI: Single Instrument stimulus
CI: Competing Instruments stimulus
Vibratos
✓ Detect vibratos in the attended track
✗ Ignore vibratos in the competing track

Figure 1. Schematic representation of the experiment. Volunteers were presented with continuous classical music pieces, Bach's Two-Part Inventions, that consisted of either a single melodic line (SI) or of two melodic lines (CI). Each melodic line was played either by a guitar or by a piano. In the CI stimuli, each melodic line was played by a different instrument. The average duration of each invention was 1.6 min; only a short segment of each invention is displayed here for clarity. Vibratos were inserted into the acoustic waveforms of each melody (gray shading). In the CI condition, the participants had to attend to one of the two instruments and identify the corresponding vibratos (green tick marks) while ignoring the other instrument and its vibratos (red crosses). The stimuli were presented in blocks composed of an SI stimulus followed by a CI stimulus during which the participant was asked to attend to the instrument that they heard before in the SI stimulus. The attended instrument was alternated between blocks, and each block was played twice such that the attended instrument differed in the two presentations. The volunteers' neural responses were recorded throughout the experiment through bipolar two-channel EEG recordings.

Two-Part inventions was presented twice: once with the SI stimulus played by the guitar and once with the SI stimulus played by the piano. Each participant therefore heard seven SI stimuli played by the guitar, seven SI stimuli played by the piano, and 14 CI stimuli.

All participants were initially presented with the same two training blocks, one with an SI stimulus played by the guitar and one played by the piano, that corresponded to the same invention. These stimuli presentations served to familiarize the participants with the task of attending to one melodic line in the CI stimulus and to identify the embedded vibratos. These training blocks were excluded from further analysis, leaving six inventions in each condition.

The presentation order of the remaining blocks was pseudorandomized across participants. In the second presentation of a given CI stimulus, a participant was asked to attend to the instrument they ignored in the first presentation. Two consecutive blocks did not correspond to the same invention. Each participant therefore heard each CI stimulus twice but attended a different instrument in each presentation. Whether a participant was initially asked to attend to the guitar or the piano was randomly decided. The participant's neural responses were measured through scalp EEG with a two-channel bipolar montage (head vertex minus mastoids).

We used encoding and decoding approaches (linear forward and backward models) to relate the acoustic stimuli to the recorded neural data. We specifically investigated the neural representation of the temporal fine structure by using the stimulus waveform as a feature. We first established that we could indeed record a significant neural response to this feature by comparing the neural responses to a null distribution at the level of individual participants as well as on the population level. We then studied the time course of the response in the region between 0 and 45 msec, using both forward and backward models. Finally, we investigated a putative attentional modulation of this neural response through contrasting the encoding of each instrument in the neural data when attended versus ignored. We used conservative filters to reduce distortions to the neural responses and their latencies but verified that our results and, in particular, the ones related to attention did not change with stronger filtering (data not shown).

Code and Data Availability

The analysis presented in this article was implemented using MATLAB (R2019b, The MathWorks, Inc.) with the EEGLAB toolbox (Delorme & Makeig, 2004). The linear forward and backward models were trained using the LMpackage (github.com/octaveEtard/LMpackage). The raw data as well as analysis code and processed data required to reproduce the results presented here have been made available (https://github.com/octaveEtard/EEGmusic2020; https://zenodo.org/record/4470135).

Participants

Seventeen volunteers (aged 23.8 ± 2.9 years, nine women) participated in this experiment. The number of participants was chosen based on previous studies investigating similar neural responses to continuous speech (Etard et al., 2019; Forte et al., 2017). All participants were right-handed, had no history of auditory or neurological impairments, and provided written informed consent. The experimental procedures were approved by the Imperial College Research Ethics Committee.

Music Stimuli

To generate neural responses to each instrument that were of similar magnitude, the notes of the guitar were lowered by one octave so that their fundamental frequencies fell below 500 Hz. They remained nonetheless somewhat higher than those of the piano notes (Figure 2A, B). The music stimuli were synthesized from MIDI files to generate wav files. These were then processed using MATLAB to apply vibratos to 10 segments in each melodic line. Each vibrato was constructed by introducing a sinusoidal warp at a modulation frequency of $f_{\mathrm{m}} = 8$ Hz on the waveform of a single note. The onset and offset times of the notes were obtained from the MIDI files using the Miditoolbox for MATLAB (Eerola & Toiviainen, 2004). The notes were selected such that the onsets of any two vibratos in a given piece, whether both played by the same or different instruments, were separated by at least 1 sec. In the six inventions used in the competing condition, there were thus 60 vibratos for each instrument in total. Overall, there were six piano vibratos during a silence on the guitar track and four guitar vibratos during a silence on the piano track. All other vibratos happened while a note was being played by the other instrument.

The waveforms of the CI stimuli, w_{mixed}, were constructed by normalizing and mixing the waveform w_{g} of the guitar and the waveform w_{p} of the piano according to their root-mean-square values (RMS): $w_{\mathrm{mixed}} = \frac{w_{\mathrm{g}}}{\mathrm{RMS}(w_{\mathrm{g}})} + 1.25 \times \frac{w_{\mathrm{p}}}{\mathrm{RMS}(w_{\mathrm{p}})}$. The mixing parameter of 1.25 for the piano was chosen following a preliminary pilot study used to ascertain the feasibility of the task and tune its difficulty, and that was conducted before the main investigation. In this independent test, the participants reported that the piano sounded weaker when the intensities of the two instruments were balanced. The mixing factor was progressively changed until they reported that the two instruments sounded equally loud and that it was as challenging to attend to either instrument. Their performance on the vibrato identification task also reflected the approximately similar difficulty in attending to either instrument.

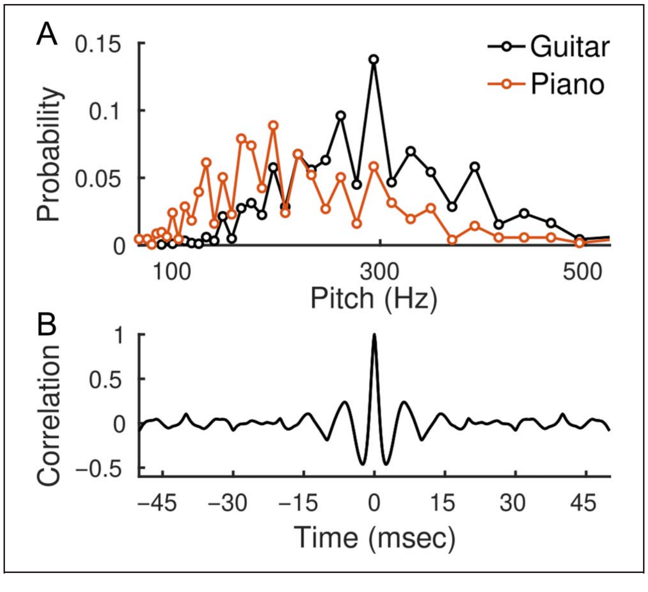

Figure 2. Properties of the acoustic stimuli and of the filtered EEG data. (A) The probability mass function of the fundamental frequency of the notes peaked at around 196 Hz for the piano (red) and at about 294 Hz for the guitar (black). Most fundamental frequencies lied between 100 and 400 Hz, with the distribution of the guitar notes being shifted to somewhat higher frequencies. (B) To eliminate frequencies below the range of the fundamental frequencies, the EEG data were high-pass filtered above 130 Hz. The filtered EEG data consequently displayed some periodicity and correlation in time as evident from its autocorrelation function.

The duration of the seven Two-Part Inventions was, taken together, 11.2 min. In the SI conditions, only the first half of the corresponding invention was played.

Behavioral Task

In the CI condition, the participants were instructed to attentively listen to one instrument while ignoring the other. They were also asked to classify the vibratos they heard by pressing a key to indicate the ones that belonged to the attended instrument. A key press within 2 sec after the onset of a vibrato in the attended or ignored instrument was classified respectively as "true positive" or "false positive." Key presses outside these ranges were classified as "unprompted" and were not analyzed further. On average, 0.5 ± 0.9 vibratos were classified as "umprompted" per stimulus and not analyzed further (average ± standard deviation over participants and stimuli). These include responses outside the time window mentioned above, as well as repeated key presses beyond the first one within a response window. Because of a technical error, behavioral data were not recorded for one participant, and only the results for the 16 remaining participants were analyzed. The sensitivity index d' was computed for each participant when attending to the guitar and the piano, and it was compared between the two conditions at the population level using a two-tailed paired Wilcoxon signed-rank test. Moreover, for each condition, the true positive rate (TPR) was compared with the false positive rate (FPR), and the TPR and FPR were compared between conditions at the population level using two-tailed paired Wilcoxon signed-rank tests with false discovery rate (FDR) correction for multiple comparisons (four tests).

Neural Data Acquisition and Stimulus Presentation

Scalp EEG was recorded through five passive Ag/AgCl electrodes (Multitrode, BrainProducts). Two electrodes were positioned close together near the cranial vertex (Cz), and two electrodes were placed on the left and right mastoid processes. A ground electrode was placed on the forehead. The impedance between each electrode and the skin was reduced below 5 kΩ using abrasive electrolyte gel (Abralyt HiCl, Easycap). One vertex electrode was paired with the left mastoid electrode, and they were connected to, respectively, the noninverting and inverting ports of a bipolar amplifier (EP-PreAmp, BrainProducts). The remaining vertex and mastoid electrodes were similarly connected to a second identical amplifier. The output of each bipolar preamplifier was fed into an amplifier (actiCHamp, BrainProducts) and digitized with a sampling frequency of 5 kHz, thus yielding two electrophysiological data channels. The audio stimuli were simultaneously recorded at 5 kHz by the amplifier through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts). This channel and independent analogue triggers delivered through an LPT port were used to temporally align the EEG data and stimuli through cross-correlation. The stimuli were delivered diotically at a comfortable loudness level through insert tube earphones (ER-3C, Etymotic) to minimize stimulation artifacts. These earphones introduced a 1-msec delay that was compensated for by shifting the neural data forward in time by 1 msec.

EEG Data Filtering

To analyze the neural responses to the temporal fine structure of the stimuli, the EEG data were high-pass filtered above 130 Hz (windowed-sinc filters, Kaiser window, one pass forward and compensated for delay; cutoff = 115 Hz, transition bandwidth = 30 Hz, order = 536). These filters rejected lower-frequency neural activity but reduced the temporal precision of the data, as evidenced by the autocorrelation function of the filtered EEG data (Figure 2C). Notably, they were noncausal filters that spread responses in both temporal directions.

The EEG data were also analyzed regarding cortical responses. To this end, it was band-pass filtered between 1 and 20 Hz and resampled to 100 Hz (windowed-sinc filters, Kaiser window, one pass forward and compensated for delay; low-pass filter: cutoff = 35 Hz, transition bandwidth = 30 Hz, order = 536; high-pass filter: cutoff = 0.5 Hz, transition bandwidth = 1 Hz, order = 322).

Stimulus Representations

The vibratos might lead to neural responses deviating from the ones elicited by the rest of the tracks. The corresponding parts of the stimulus waveforms were thus replaced with zeros to create the stimulus representations (features) used in the encoding and decoding models. These waveforms were then low-pass filtered and resampled from 44.1 to 5 kHz, the sampling frequency of the corresponding EEG data, using a linear phase finite impulse response antialiasing filter (windowed-sinc filter, Kaiser window, one pass forward and compensated for delay; cutoff = 2250 Hz, transition bandwidth = 500 Hz, order = 14,126).

To derive cortical responses to the stimuli, features encoding the note onsets were created as well. These were formed by time series with a constant value of equal to 0, except at note onsets, which were marked by an impulse with a value of 1. These were created based on the timing of the note onset as extracted from the MIDI files, at a sampling frequency matching the EEG data for cortical analysis (100 Hz).

Encoding Models

We used regularized linear forward models to derive the neural response to the stimulus waveform. In these convolutive encoding models, the measured EEG response e is modeled as $e(t) = (r * s)(t) = n(t)$, where s is the stimulus waveform, r is the neural response or temporal response function (TRF), n is noise, and * is the convolution symbol. In practice, assuming a nonzero response in a time interval ($\tau_{\min}$, $\tau_{\max}$) only and with discrete data, the EEG activity $e_i(t_n)$ at channel $i \in \{1, 2\}$ and at time t_n can be estimated as $\hat{e}_i(t_n) = \sum_{k=1}^{N} r(\tau_k) \times s(t_n - \tau_k)$, with $\tau_1 = \tau_{\min}$ and $\tau_N = \tau_{\max}$. Given the bipolar montage we used, as well as the diotic stimulus presentation, we did not expect any difference between the two EEG channels and assumed the same neural response for both.

The model was estimated for time lags spanning $\tau_{\min} = -100$ msec to $\tau_{\max} = 45$ msec. A population-averaged TRF r was fitted using ridge regression coupled with a leave-one-participant-out and leave-one-data-part out cross-validation (Crosse, Di Liberto, Bednar, & Lalor, 2016; Hastie, Tibshirani, & Friedman, 2009; Lalor, Power, Reilly, & Foxe, 2009). In details, the neural responses for one participant during one stimulus part were chosen as testing data. The model coefficients were derived using the neural data from all the other participants, in response to all the other stimulus parts. The model was evaluated on the testing data that were hence not seen by the model during training. This constituted one cross-validation fold. The left-out participant and the left-out data part were then iterated until all combinations were exhausted, for a total of 17 × 6 = 102 folds. The validation performance of the model was quantified by dividing the predicted neural response $\hat{e}_i$ and the measured EEG activity e_i from the testing data in each fold into 10-sec long segments and by computing Pearson's correlation coefficients between each segment. The correlation coefficients thus obtained were then averaged over all cross-validation folds as well as over all EEG channels.

The performance was assessed for models corresponding to 25 normalized regularization coefficients λ_n that were distributed uniformly on a logarithmic scale between 10^{-6} and 10^{6}. The regularization coefficient was thereby $\lambda = \lambda_n \times m$, with m as the mean eigenvalue of the predictor's autocorrelation matrix (Biesmans, Das, Francart, & Bertrand, 2017). Regularization is used in the EEG literature in conjunction with linear convolutive models to penalize large, oscillating coefficients and prevent overfitting, thus increasing the models' generalizability and predictive performances (Wong et al., 2018;

Biesmans et al., 2017). The model yielding the highest reconstruction performance was chosen as representing the neural response. To assess the significance of the obtained TRFs, the negative, noncausal part of the response, −100 to 0 msec, was used to construct a null distribution. For each instrument, a Gaussian distribution was fitted to the pooled data points from the negative part of the response. From the distribution, we determined the p values of all the points in the positive part of the response (0–45 msec) and applied an FDR correction for multiple comparison over time points and instruments.

To ascertain the relative contributions of the onset and of the sustained parts of the notes to the neural response, we created a new representation of the stimuli in which the note onsets were suppressed. This was achieved by multiplying the original stimulus waveforms by a 60-msec window w centered on each note onsets, with $w(t) = 1 - h(t)$ and h representing a 60-msec Hann window. Forward models were then derived for the original stimuli, and their onset-suppressed versions for the two SI conditions were taken together by pooling the data from both instruments. These two models were fitted, and their significance was ascertained as described above, that is, by comparing the causal part to the null models, with FDR correction for multiple comparisons over time points and over the two models. In the cross-validation procedure, two data parts, one from each SI condition and corresponding to the same invention, were left out at each stage.

Decoding Models

We also used backward models to reconstruct the stimulus waveform s as a linear combination of the neural activity e_i on each channel i at different time lags: $\hat{s}(t_n) = \sum_{i=1}^{2} \sum_{k=1}^{N} \beta_i(\tau_k) \times e_i(t_n + \tau_k)$, with $\tau_{\min} \leq \tau_k \leq \tau_{\max}$. The coefficients β were trained for each participant independently, using ridge regression with a leave-one-part-out cross-validation and a normalized regularization coefficient $\lambda_n = 10^{-0.5}$ (Biesmans et al., 2017). As with the forward models, the performances of the backward models were measured through computing the correlation coefficients between the reconstructed stimulus and the actual one on 10-sec long segments of the testing data. The set of correlation coefficients pooled from all cross-validation folds for a given participant was used when performing statistical testing at the level of individual participants, and the corresponding average correlation coefficient was used for each participant when testing at the population level. The performances of the models were thereafter used to quantify the neural encoding of each stimulus for a given reconstruction time window $\tau_{\min} - \tau_{\max}$.

Significance of the Stimulus Reconstruction

The neural encoding of the SI stimuli for each instrument was measured through the backward models using reconstruction time windows of equal duration but centered on different delays. To establish the significance of the stimulus reconstruction procedure at the level of individual participants, a window of delays between $\tau_{\min} = -15$ msec and $\tau_{\max} = 0$ msec was used to provide a null distribution for each participant. The neural encoding in the window of interest, from $\tau_{\min} = 0$ msec to $\tau_{\max} = 15$ msec, was compared with the null distribution for each participant using one-tailed paired Wilcoxon signed-rank tests with FDR correction for multiple comparisons over participants and instruments. Significance was also derived at the population level using the mean correlation coefficients for each participant from the null window of negative delays to create a null population-level distribution. To test the time windows in which a significant response could be detected, the mean reconstruction accuracies from three windows of interest (0–15, 15–30, and 30–45 msec) were compared with this null distribution using one-tailed paired Wilcoxon signed-rank tests with FDR correction for multiple comparisons over windows and instruments.

Because the guitar and piano waveforms formed pairs derived from the same inventions and although their frequency contents were different, one may wonder whether one instrument could be predicted from the other and, in turn, whether the neural responses to one instrument could be predicted or used to decode the other one. To address this question, we trained linear backward models that sought to reconstruct the waveform of one instrument from the neural data that were recorded when the other instrument from the same invention was played in the SI conditions (0–15 msec reconstruction window). The model performance was then compared with the null distribution previously described (obtained from a −15 to 0 msec reconstruction window) at the population level, using one-tailed paired Wilcoxon signed-rank tests.

Competing Conditions, Attended and Ignored Instruments

In the CI conditions, we trained backward models to reconstruct the waveform of either the attended or the ignored instrument independently, using a window of temporal delays from $\tau_{\min} = 0$ msec to $\tau_{\max} = 15$ msec as detailed above. We then compared the neural encoding of each instrument, when attended and when ignored, at the population level using two-tailed paired Wilcoxon signed-rank tests with FDR correction for multiple comparisons over instruments. Furthermore, to ascertain the significance of the reconstructions, noise models were similarly trained for both instrument and attention condition using a time window from $\tau_{\min} = -15$ msec to $\tau_{\max} = 0$ msec. The performance of the meaningful models was then compared with these null distributions at the population level, using one-tailed paired Wilcoxon signed-rank tests with FDR correction for multiple comparisons over instruments and attention conditions.

We also used forward models reconstructing the neural activity as the sum of two neural responses, one to the attended instrument and one to the ignored one. In this instance, the EEG response e is modeled as $e(t) = (r_A * s_A)(t) + (r_I * s_I)(t) + n(t)$, where s_A and s_I are the attended and ignored stimulus waveforms and r_A and r_I are the corresponding TRFs. In a similar manner to the procedures previously described, population-averaged TRFs were fitted using ridge regression coupled with a leave-one-participant-out and leave-one-data-part out cross-validation for time lags spanning $\tau_{min} = -100$ msec to $\tau_{max} = 45$ msec on the pooled data from the two CI conditions. To assess the presence of a putative attentional modulation in the obtained TRFs, the distribution of amplitude across participants was compared between the attended and ignored TRFs for each time point in the 0–15 msec ROI (two-tailed paired Wilcoxon signed-rank tests with significance threshold $p \leq .01$ after FDR correction for multiple comparisons over time points).

To investigate a putative attentional modulation in cortical responses, the same procedure was repeated using the EEG data, band-pass filtered between 1 and 20 Hz and the stimulus feature marking the note onset with an impulse. In this case, the models were fitted with time lags between $\tau_{min} = -250$ msec to $\tau_{max} = 750$ msec. Previous continuous speech studies revealed attentional modulation of the neural responses at latencies around 100 msec (N100; Ding & Simon, 2012a). An ROI from 50 to 200 msec was thus defined, and the distribution of amplitude across participants was compared between the attended and ignored TRFs for each time point in this ROI (two-tailed paired Wilcoxon signed-rank tests with tests with significance threshold $p \leq .01$ after FDR correction for multiple comparisons over time points).

RESULTS

We asked volunteers to attend to continuous musical pieces consisting of either one single instrument (SI) or of two competing instruments (CI) while we recorded their neural activity using EEG (Figure 1). We first sought to analyze the neural response to the temporal fine structure of a single melodic line. To this end, we computed a linear forward model to derive neural responses to the stimulus waveform at the population level in the SI conditions (Figure 3A). The TRFs that we obtained for the two instruments were qualitatively similar to each other. They displayed a major significant response at a latency of 7.6 msec, as well as a minor positive peak at 2.2 msec, with side lobes reminiscent of the EEG autocorrelation function (Figure 2C).

The neural response to temporal fine structure may be related to the well-established frequency-following response (FFR). Because the latter is known to first exhibit a response to a stimulus onset and to then follow the sustained features, we explored the relative contributions of the note onsets and their sustained oscillations to the neural response. We therefore trained a forward model with stimulus waveforms in which the note onsets were suppressed (Figure 3B). The obtained TRFs had similar significant regions and resembled the TRFs to the original stimulus waveforms. Moreover, the causal parts of the two TRFs, those with positive delays, were highly correlated ($r = .96$).

As an alternative method to the forward models, we then also used decoding models that reconstructed the stimulus waveforms based on the EEG data. We computed these models for each participant in the SI condition. To ascertain the statistical significance of the reconstructions, we used a window from −15 to 0 msec to provide a null distribution of performance. Compared with this chance level, we found that a significant reconstruction accuracy could be obtained for most participants when using

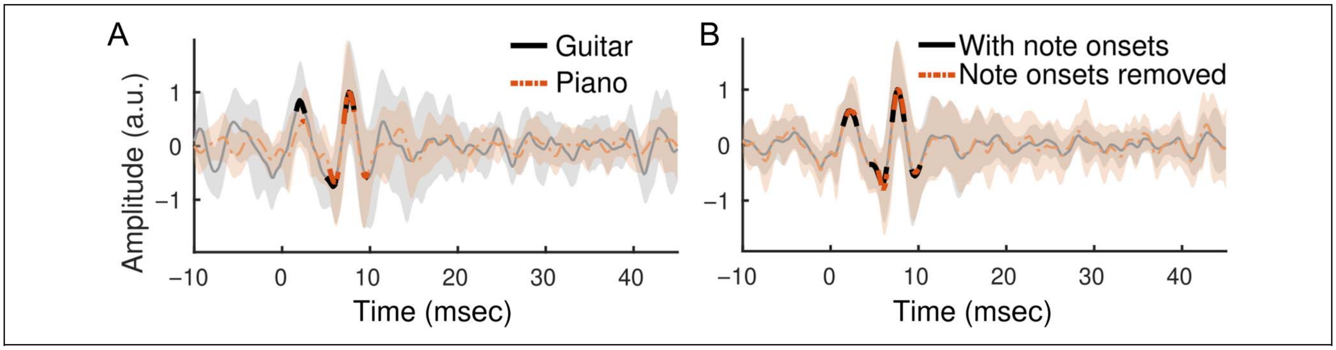

Figure 3. TRFs. (A) We obtained TRFs on the population level from forward models that predicted the neural responses from the stimulus temporal fine structure in the SI conditions for the guitar (black) and for the piano (red). Shaded regions denote ±1 standard deviation across participants around the mean TRFs. Significant regions (thick lines) emerged at similar latencies for the guitar and the piano, with a first peak at 2.2 msec, followed by a main positive peak at 7.6 msec. (B) We also computed TRFs for both instruments taken together from stimulus waveforms in which the note onsets where removed ("Note onsets removed," red). The obtained TRFs exhibited nonetheless the same significant peaks as the TRFs from the original temporal fine structure feature ("With note onsets," black), indicating that the neural response was not influenced by the note onsets. a.u. = arbitrary units.

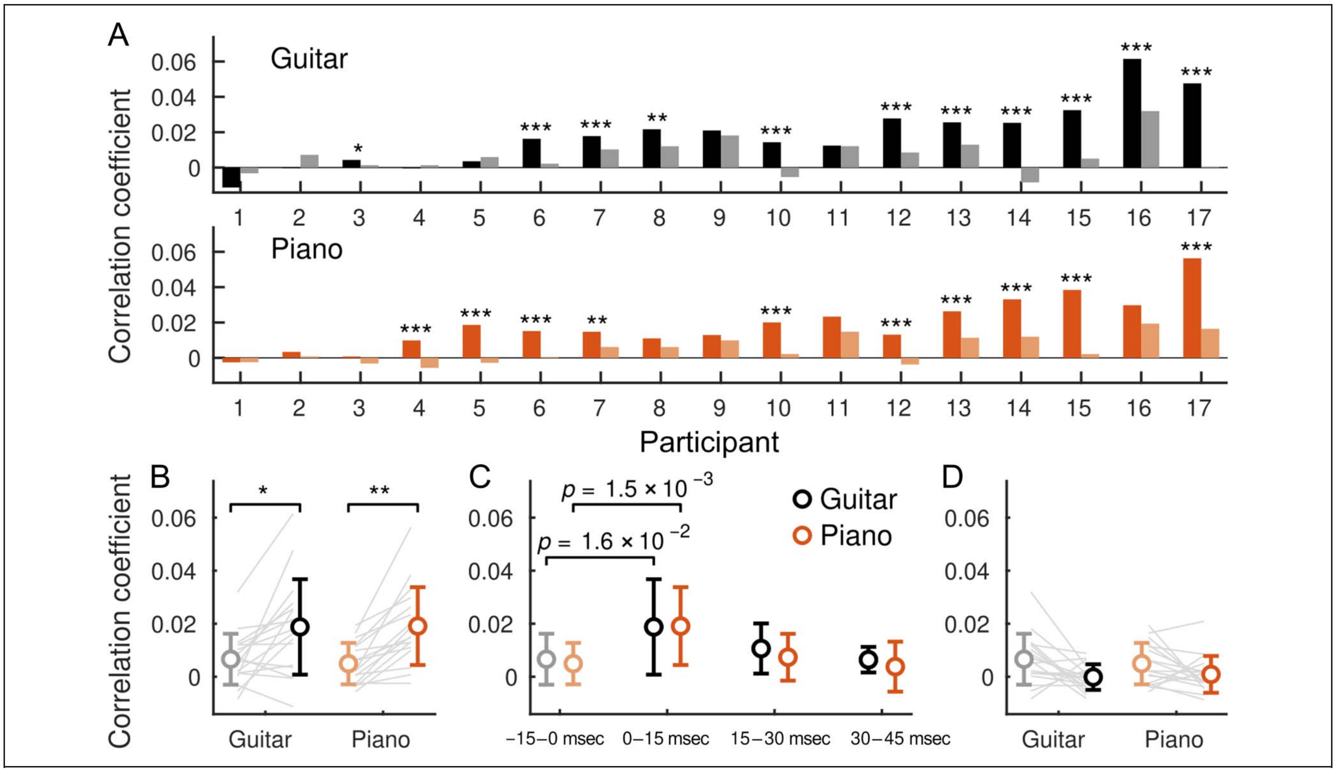

Figure 4. Backward models that reconstruct the stimulus waveform from the EEG data in the SI condition. (A) In most participants, the backward models gave a stimulus reconstruction that had a significantly larger correlation (dark color) with the original waveform than a null model (light color). The volunteers were sorted by mean performance, and asterisks indicate p values (*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$). (B) The mean reconstruction accuracy for each participant was used to test the significance of the reconstruction at the population level. Both the guitar and piano stimuli could be reconstructed significantly better from the EEG recordings than from null models. (C) We also assessed the reconstruction of the backward models using three windows of temporal delays of 0–15, 15–30, and 30–45 msec (dark colors) and compared them to a null model obtained from the negative delays of −15 to 0 msec (light colors). Only the temporal window of 0–15 msec allowed for a stimulus reconstruction that was significantly better than that of the null model. (D) Reconstructing one instrument waveform using the EEG recorded during the presentation of the other instrument (0–15 msec window; dark colors) did not yield significant performances as compared with the null model derived by using negative delays (−15 to 0 msec; light colors). Although the two instrument waveforms formed pairs corresponding to an invention, the waveform of one instrument could not be predicted from the neural responses to the other instrument.

time lags from 0 to 15 msec for both guitar and piano (Figure 4A). Indeed, significant reconstructions of the guitar waveforms were obtained in 11 out of 17 participants ($p \leq .05$), in 10 participants for the piano waveforms, and in 8 participants for both types of stimuli. The reconstructions of the waveforms for the guitar and for the piano were also significant at the population level (Figure 4B; guitar: $p = 1.6 \times 10^{-2}$; piano: $p = 1.54 \times 10^{-3}$). Finally, on the population level, when assessing the statistical significance of the stimulus reconstructions using each of three windows of interest (0–15, 15–30, and 30–45 msec), we found that only the window from 0 to 15 msec yielded a significant reconstruction accuracy, for either instrument (Figure 4C).

As the stimuli we used were derived from the left and right hands of inventions, one may wonder whether two instrument waveforms derived from the same piece are independent and whether the neural responses to one instrument could be used to decode the other one. This is particularly relevant in the context of the attention experiment where such an effect could obscure a putative attentional modulation. However, the stimulus reconstruction accuracy when mismatching the EEG–stimuli pairs in such a way (0–15 msec reconstruction window) was not significant as compared with the null distribution using matched EEG–stimuli pairs and a −15 to 0 msec reconstruction window (Figure 4D; guitar: $p = .99$, piano: $p = .96$).

Armed with the ability to measure neural responses to the temporal fine structure of the notes in a particular melody, we then investigated whether this response was affected by selective attention. To this end, we analyzed the CI stimuli, in which the participants had to attend selectively to one instrument while ignoring the other. We monitored attention by asking the volunteers to classify vibratos inserted into the melodic line played by the target instrument. The participants exhibited varied performances on this task; however, they all had an average performance that was better than that of a random observer, as shown by their receiver operating characteristics, when selectively attending to either of the two instruments (Figure 5A). Accordingly, at the population

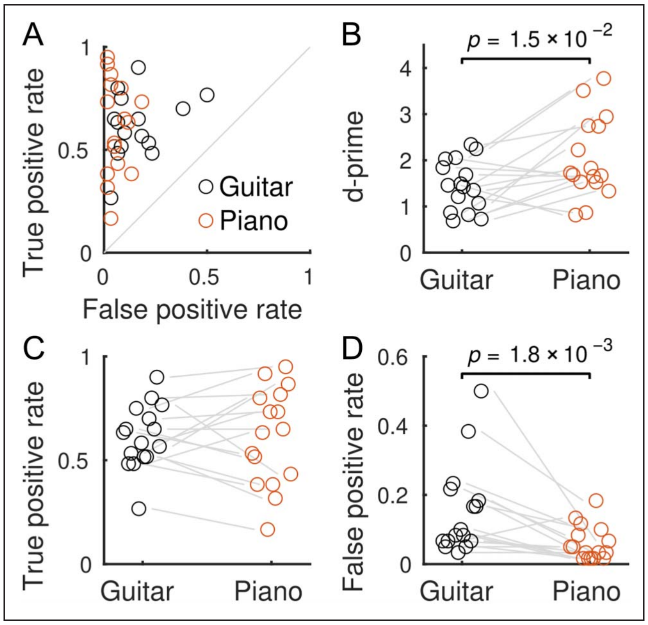

Figure 5. Behavioral results for the vibrato classification, task. Each circle represents a participant. (A) The receiver operating characteristics show that each participant performed above chance level in the CI condition, both when attending to the guitar (black) and when attending to the piano (red). (B) The average sensitivity index d' was significantly larger when attending to the piano than to the guitar ($p = 1.5 \times 10^{-2}$) with an average value of 2.0 and 1.5, respectively. (C) The rate of true positives was similar when attending to the guitar and then attending to the piano. (D) Attending to the guitar led to more false positives than attending to the piano ($p = 1.8 \times 10^{-3}$).

level, the TPR was significantly larger than the FPR when attending to either instrument ($p < 10^{-3}$ for guitar and piano). The sensitivity index d' was significantly larger when attending to the piano than when attending to the guitar ($p = 1.5 \times 10^{-2}$), with an average value of 2.0 and 1.5, respectively (Figure 5B). The TPR did not differ significantly between the two CI conditions ($p = .88$; Figure 5C), but the FPR was significantly higher when the participants were attending to the guitar compared with the piano (FPR: $p = 1.79 \times 10^{-3}$; Figure 5D).

To test for a putative attentional modulation of the encoding of the stimulus temporal fine structure, we first used backward models with a window from 0 to 15 msec to reconstruct each instrument waveform when it was attended as well as when it was ignored. The reconstruction accuracies did, however, not exhibit a statistically significant difference between the attended and the ignored case (Figure 6A; $p = .49$ for guitar and piano). To ascertain the statistical significance of the reconstructions, we used a window from −15 to 0 msec to provide a null distribution of performance for each instrument and attention condition. Compared with this chance level, we found that the piano and guitar, when they were attended and ignored, yielded significant reconstructions (Figure 7A; guitar attended: $p = 6 \times 10^{-4}$, guitar ignored: $p = .013$, piano attended: $p = 6 \times 10^{-4}$, piano ignored $p = 6 \times 10^{-4}$).

We then computed a linear forward model that included two features, the attended and ignored instruments. The linear forward model was trained using the pooled data from the two CI conditions. The model then allowed us to compare the amplitude of the attended and ignored TRFs at each time lag from 0 to 15 msec. No significant difference between the amplitudes emerged at any temporal lag (Figure 6B). Finally, linear forward models reconstructing cortical neural activity based on the note onsets for the attended and ignored instruments were trained using the pooled data from the two CI conditions. The amplitude of the attended and ignored TRFs at each time lag was then compared in an ROI from 50 to 200 msec. The amplitude of the attended and ignored TRFs were found to significantly differ at time lags from 100 to 160 msec (Figure 7B, $p \leq .01$).

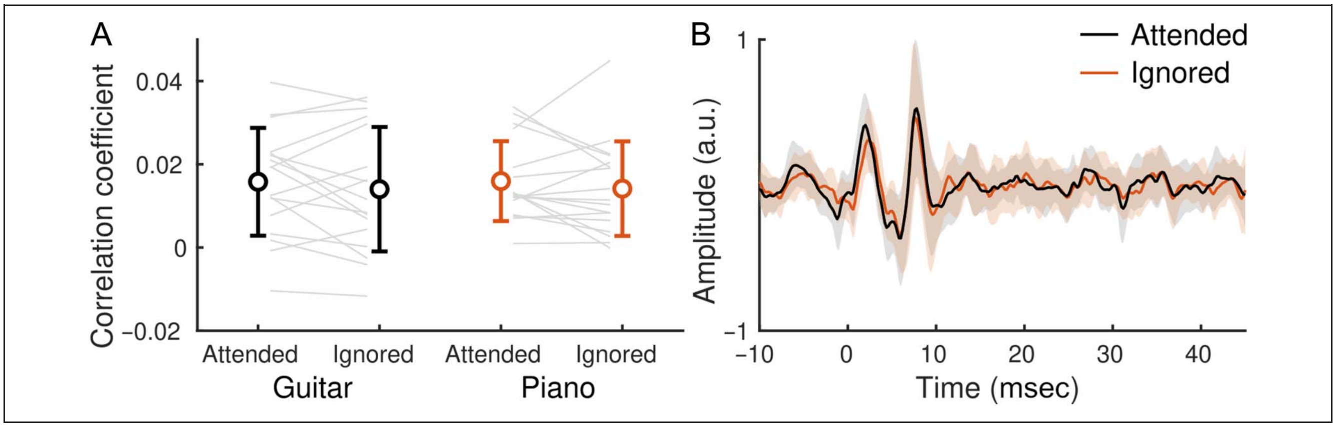

Figure 6. Absence of attentional modulation of subcortical neural responses. (A) Backward models were trained to reconstruct the stimulus waveforms for the guitar (black) and piano (red) in the CI conditions when they were attended or ignored. The reconstruction accuracies, as assessed by the correlation coefficient between the reconstructed and the original signals, did not differ significantly between the attended versus the ignored cases ($p = .49$ for guitar and piano). (B) Population-averaged TRFs were derived over the two CI conditions taken together for the attended (black) and ignored (red) instruments. The amplitude of the obtained TRFs did not significantly differ in the 0–15 msec ROI. Shaded regions denote ±1 standard deviation across participants around the mean TRFs. a.u. = arbitrary units.

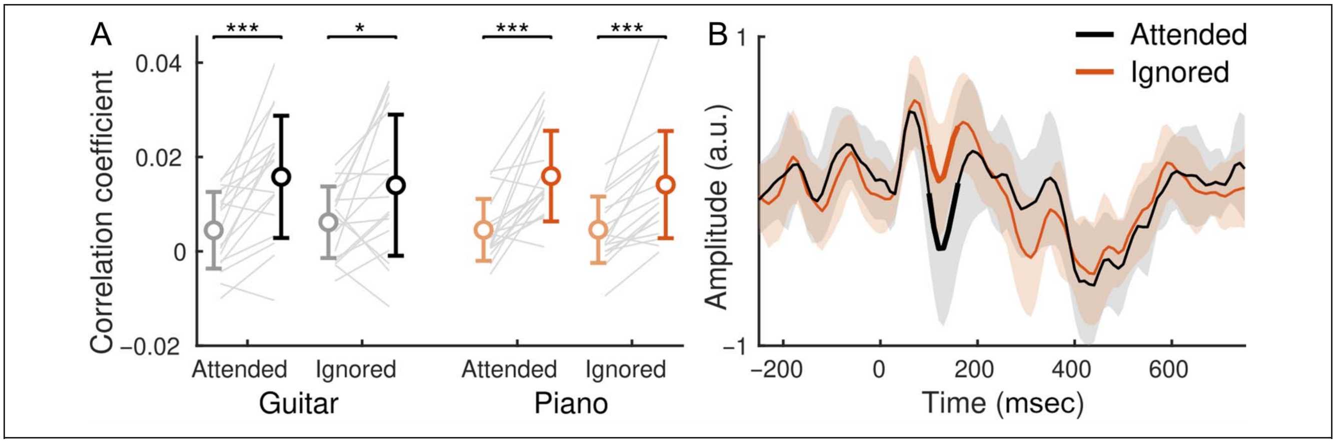

Figure 7. Additional controls and cortical attentional modulation. (A) Comparing the performance of the backward models in the CI condition to null distributions revealed that they performed significantly better than chance, for both instrument and attention condition (guitar attended: $p = 6.1 \times 10^{-4}$, guitar ignored: $p = .013$, piano attended: $p = 6.1 \times 10^{-4}$, piano ignored: $p = 6.1 \times 10^{-4}$). (B) Training linear forward models to reconstruct slow cortical neural activity (1–20 Hz) based on note onsets showed that population-averaged TRF exhibited attentional modulation in the 100–160 msec region. a.u. = arbitrary units.

DISCUSSION

We showed for the first time that neural responses to the temporal fine structure of continuous musical melodies can be obtained from EEG recordings using linear convolutive models. In particular, we demonstrated that the EEG recordings could in part be predicted from the acoustic waveforms (forward model; Figure 3). Vice versa, the temporal fine structure of the musical stimuli could be decoded from the corresponding EEG recordings (backward model; Figure 4). Significant responses could be obtained in most individual participants when they were exposed to about 5 min of a single melodic line.

The neural response at the population level revealed further information about its origin. Indeed, the significant parts of the response, as obtained from the forward models, emerged most strongly at the latency of 7.6 msec (Figure 3A). The responses at the other latencies may have reflected our use of high-pass filters for the EEG data, which spread the response in time in both directions (Widmann, Schröger, & Maess, 2015). The autocorrelation of the filtered EEG data exhibited side lobes that are reminiscent of the structure of some of the peaks that we obtained in the neural responses (Figure 2C).

The backward model showed likewise that only delays between 0 and 15 msec allowed for a significant reconstruction of the stimulus waveform. Together with the evidence from the forward model, these delays suggest a subcortical origin of the neural response, putatively in the inferior colliculus, although different subcortical structures may contribute as well (Bidelman, 2015, 2018; Skoe & Kraus, 2010; Sohmer, Pratt, & Kinarti, 1977). Recent MEG work indeed uncovered cortical contributions to the FFR in humans (Ross, Tremblay, & Alain, 2020; Hartmann & Weisz, 2019; Coffey et al., 2016), although they may be limited to frequencies below 150 Hz (Bidelman, 2018). The scalp-recorded FFR may accordingly combine multiple subcortical and cortical sources (Coffey et al., 2019). Although the neural response that we have described here is arguably of subcortical origin, our use of only two EEG channels may have obstructed the observation of later cortical sources with different dipole orientations.

Neural responses can occur to both transient (e.g., clicks, onsets) and sustained features of complex stimuli. When investigating the FFR, for instance, these two aspects can be segregated by time regions (Skoe & Kraus, 2010). However, the continuous nature of the stimuli that we used here did not allow for this type of analysis. Instead, we trained a forward model with stimulus waveforms where note onsets were suppressed and compared it to a forward model trained using the intact waveforms (Figure 3A, B). The two responses were strikingly similar, suggesting that they are primarily driven by the sustained periodic oscillations of individual notes rather than their onsets. This may be expected, as these sustained oscillations accounted for most of our music stimuli. In a click train, in contrast, the stimuli are entirely constituted of transients.

When the participants were presented with stimuli consisting of two competing instruments, they had to selectively attend to one of them and identify vibratos that were inserted in the melodic line of that instrument. We used this task as a marker of selective attention, comparable to the use of comprehension questions in the case of speech stimuli. Correctly attributing the vibratos to the appropriate stream (i.e., reacting only to the ones in the attended instrument) required attentively monitoring the melodies. Each competing piece contained 20 vibratos in total (10 per instruments) for a duration of a few minutes, thus yielding a vibrato every few seconds and encouraging continuous engagement. We found that most participants were able to identify the target vibratos while ignoring the distractors (Figure 5). The sensitivity index d' was significantly larger when attending to the piano than the guitar. When attending to either instrument, the TPR did not significantly differ, but the FPR was lower when attending to the piano, indicating that this effect mediated the difference in d' values. The instrument mixing ratio used in the current study was determined following a small pilot study to even out perceptual intensity and task difficulty (Methods section), but it may have contributed to this effect. We also hypothesize that because pianos cannot naturally produce vibratos, the participants may have had a bias leading to a higher propensity to attribute vibratos to the guitar (Figure 5D).

The task of attending to one of two melodic lines allowed us to investigate whether the neural response to the temporal fine structure of a particular melodic line was modulated by selective attention. Following our results on the statistical methods for obtaining this neural response, we employed backward models to reconstruct the stimulus waveform from the EEG recording, using temporal delays between 0 and 15 msec. We did not, however, find any significant difference between the resulting reconstruction accuracies of a melodic line when it was being attended or ignored for either instrument (Figure 6A). To verify this result using a different methodological approach, we also trained a forward model that used the attended and ignored instruments as features. Comparing the amplitude of the attended and ignored TRFs between 0 and 15 msec did not reveal any significant difference (Figure 6B).

Backward models have typically smaller reconstruction accuracies in competing conditions than in response to isolated targets. This has not prevented previous work from observing attentional modulation (Etard et al., 2019; Forte et al., 2017; O'Sullivan et al., 2015) but is problematic if the observed model performances are not above chance level. To ensure that the reconstructions we obtained were indeed relevant, we confirmed that they were significantly greater than null distributions for both instrument and attention condition (Figure 7A).

Finally, we trained forward models to reconstruct cortical neural activity using features encoding the note onsets of the attended and ignored instruments. Comparing the amplitude of the obtained TRFs in an ROI from 50 to 200 msec, we found that a significant difference emerged from 100 to 160 msec (Figure 7B), in a similar manner to the attentional modulation observed in continuous speech (N100; Ding & Simon, 2012a). This indicates that, although we could not observe any modulation in the subcortical responses, the low-frequency cortical neural activity did exhibit an attentional modulation consistent with the behavioral task.

Our negative finding regarding attentional modulation contrasts with previous work on similar neural responses to the temporal fine structure of speech that were found to be modulated by selective attention (Etard et al., 2019; Forte et al., 2017). It also contrasts with recent MEG work that showed that the cortical components of the FFR can be modulated by intermodal attention (Hartmann & Weisz, 2019).

These differences may point to underlying differences between music and speech. First, the two melodic lines that we used in the present work may have been difficult to selectively attend, because they originated from one musical piece, were contrapuntal, and often followed or responded to each other. The resulting interaction between the two melodic lines makes their juxtaposition rather different from that of two independent competing voices that do not interact but merely generate informational and acoustical masking. Although two competing speakers may encourage selective attention and neural processing of one of them, our two melodic lines may therefore rather encourage attention, as well as neural processing, of the acoustic mixture.

Silences in the target instrument could also favor shifts of attention and potentially undermine neural markers of selective attention. However, silences in one of the two competing streams are not exclusive to music but form, for example, an intrinsic part of continuous speech as well, contributing to its rhythm by separating words, phrases, and sentences. In the current study, the stimuli were nonetheless chosen to minimize the amount of time where only one of the two instruments was playing in the competing condition to mitigate this effect. We found that the passages in which one instrument was silent accounted for 13% ± 4% of the duration of the competing stimuli (mean ± standard deviation over the six inventions), which is comparable to the duration of silences in previous speech studies. It is thus unlikely that these alone would explain the observed absence of attentional modulation.

Second, musical training was not an inclusion criterion in this study. Conversely, the subjects participating in a competing speaker experiments effectively have a lifelong training in isolating one speaker from noise because of the relevance of this task in daily life. As already hinted at above, we speculate that musical stimuli are instead generally perceived as a whole and that most our participants were unfamiliar with focusing on one of several instruments. We indeed observed important variability in the volunteer's performance during this study (Figure 5). Segregation of sources relies on acoustic cues (Moore & Gockel, 2012) but also on learnt patterns of the stimulus regularities. Yamagishi, Otsuka, Furukawa, and Kashino (2016) furthermore showed that FFR responses in a bistable "A-B-A" streaming experiment were modulated by thalamocortical activity according to the participants' perception. The acoustic cues alone and lack of experience of the participants in this study may not have allowed them to modulate their lower-level stimulus representation. The discrepancy between this study and previous speech results suggest that this modulation may be task dependent and rely on experience, rather than generic.

Musicians, in contrast, may in general be more familiar and trained at this task, and previous studies have indeed demonstrated that subcortical encoding of the temporal fine structure and FFR responses can exhibit long-term plasticity and that they can be modulated by musical experience (Kraus & White-Schwoch, 2017; Bidelman, Gandour, & Krishnan, 2011; Bidelman, Krishnan, & Gandour, 2011). Accordingly, musicians might exhibit attentional modulation of the neural response to the temporal fine structure of melodies, although people without musical training might not.

Finally, this study design was informed by published work analyzing similar neural responses to speech (Etard et al., 2019; Maddox & Lee, 2018; Forte et al., 2017). A combination of the factors listed above may have contributed to produce neural responses differing from the ones previously reported for speech stimuli and thus yielding no attentional modulation or one of a much smaller magnitude that could not be detected here. Further work is required to disentangle the potential effects of these hypotheses.

Music is a rich signal that consists of many transient and sustained features. Here, we focused on the comparatively high-frequency neural response to the temporal fine structure. Other features, however, could be studied as well from the same stimuli, including notably cortical responses to amplitude fluctuations. Similar cortical responses to continuous speech have received significant attention in the past years and have been shown to reflect attention (O'Sullivan et al., 2015; Ding & Simon, 2012a; Power, Foxe, Forde, Reilly, & Lalor, 2012) as well as semantic features (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018), surprisal (Weissbart, Kandylaki, & Reichenbach, 2020), or comprehension (Etard & Reichenbach, 2019; Kösem & van Wassenhove, 2017). It has indeed been found recently that the cortical encoding of sequences of tones in a melody reflects a listener's expectation of the upcoming notes (Di Liberto et al., 2020). Studying the interaction of such cortical responses with the subcortical activity related to the temporal fine structure that we have uncovered here may further clarify the neural mechanisms that allow us to perceive complex musical stimuli in their entirety, while also allowing us to selectively focus on a particular instrument or melodic line.

Acknowledgments

This research was supported by the Royal British Legion Centre for Blast Injury Studies and by EPSRC grants EP/M026728/1 and EP/R032602/1. We are grateful to the Telluride Neuromorphic Cognition Engineering Workshop 2018, in particular to discussions with Carol Krumhansl, Alain de Cheveigné, and Malcolm Slaney, as well as to the Imperial College High Performance Computing Service (doi: 10.14469/hpc/2232).

Reprint requests should be sent to Tobias Reichenbach, Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander University Erlangen-Nuremberg, Konrad-Zuse-Strasse 3/5, Erlangen, Germany, 91052, or via e-mail: tobias.j.reichenbach@fau.de.

Funding Information

Tobias Reichenbach, Engineering and Physical Sciences Research Council (https://dx.doi.org/10.13039/501100000266), grant numbers: EP/M026728/1, EP/R032602/1, Royal British Legion for Center for Blast Injury Studies (https://dx.doi.org/10.13039/100008631).

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*, 139–157. https://doi.org/10.1097/AUD.0b013e31816453dc, PubMed: 18595182

Bidelman, G. M. (2015). Multichannel recordings of the human brainstem frequency-following response: Scalp topography, source generators, and distinctions from the transient ABR. *Hearing Research*, *323*, 68–80. https://doi.org/10.1016/j.heares.2015.01.011, PubMed: 25660195

Bidelman, G. M. (2018). Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. *Neuroimage*, *175*, 56–69. https://doi.org/10.1016/j.neuroimage.2018.03.060, PubMed: 29604459

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience*, *23*, 425–434. https://doi.org/10.1162/jocn.2009.21362, PubMed: 19925180

Bidelman, G. M., Krishnan, A., & Gandour, J. T. (2011). Enhanced brainstem encoding predicts musicians' perceptual advantages with pitch. *European Journal of Neuroscience*, *33*, 530–538. https://doi.org/10.1111/j.1460-9568.2010.07527.x, PubMed: 21198980

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*, *25*, 402–412. https://doi.org/10.1109/TNSRE.2016.2571900, PubMed: 27244743

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT Press. https://doi.org/10.1121/1.408434

Bregman, A. S., Liao, C., & Levitan, R. (1990). Auditory grouping based on fundamental frequency and formant peak frequency. *Canadian Journal of Psychology*, *44*, 400–413. https://doi.org/10.1037/h0084255, PubMed: 2224643

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*, 803–809. https://doi.org/10.1016/J.CUB.2018.01.080, PubMed: 29478856

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*, 975–979. https://doi.org/10.1121/1.1907229

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications*, *7*, 11070. https://doi.org/10.1038/ncomms11070, PubMed: 27009409

Coffey, E. B. J., Nicol, T., White-Schwoch, T., Chandrasekaran, B., Krizman, J., Skoe, E., et al. (2019). Evolving perspectives on the sources of the frequency-following response. *Nature Communications*, *10*, 5036. https://doi.org/10.1038/s41467-019-13003-w, PubMed: 31695046

Cross, I., Hallam, S., & Thaut, M. (2008). *The Oxford handbook of music psychology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199298457.001.0001

Crosse, M. J., Di Liberto, 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. https://doi.org/10.3389/fnhum.2016.00604, PubMed: 27965557

de Cheveigné, A., Kawahara, H., Tsuzaki, M., & Aikawa, K. (1997). Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. *Journal of the Acoustical Society of America*, *101*, 2839–2847. https://doi.org/10.1121/1.418517

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009, PubMed: 15102499

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030, PubMed: 26412129

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., et al. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife*, *9*, e51784. https://doi.org/10.7554/eLife.51784, PubMed: 32122465

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 11854–11859. https://doi.org/10.1073/pnas.1205381109, PubMed: 22753470

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*, 78–89. https://doi.org/10.1152/jn.00297.2011, PubMed: 21975452

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311. https://doi.org/10.3389/fnhum.2014.00311, PubMed: 24904354

Eerola, T., & Toiviainen, P. (2004). *MIDI toolbox: MATLAB tools for music research*. Kopijyvä, Jyväskylä, Finland: University of Jyväskylä. www.jyu.fi/musica/miditoolbox/

Etard, O., Kegler, M., Braiman, C., Forte, A. E., & Reichenbach, T. (2019). Decoding of selective attention to continuous speech from the human auditory brainstem response. *Neuroimage*, *200*, 1–11. https://doi.org/10.1016/j.neuroimage.2019.06.029, PubMed: 31212098

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience*, *39*, 5750–5759. https://doi.org/10.1523/jneurosci.1828-18.2019, PubMed: 31109963

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife*, *6*, e27203. https://doi.org/10.7554/eLife.27203, PubMed: 28992445

Hartmann, T., & Weisz, N. (2019). Auditory cortical generators of the frequency following response are modulated by intermodal attention. *Neuroimage*, *203*, 116185. https://doi.org/10.1016/j.neuroimage.2019.116185, PubMed: 31520743

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/b94608

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, *17*, 1875–1902. https://doi.org/10.1162/0899766054322964, PubMed: 15992485

Kashino, M., & Kondo, H. M. (2012). Functional brain networks underlying perceptual switching: Auditory streaming and verbal transformations. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *367*, 977–987. https://doi.org/10.1098/rstb.2011.0370, PubMed: 22371619

Kösem, A., & van Wassenhove, V. (2017). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*, 536–544. https://doi.org/10.1080/23273798.2016.1238495

Kraus, N., & White-Schwoch, T. (2017). Neurobiology of everyday communication: What have we learned from music? *Neuroscientist*, *23*, 287–298. https://doi.org/10.1177/1073858416653593, PubMed: 27284021

Krizman, J., & Kraus, N. (2019). Analyzing the FFR: A tutorial for decoding the richness of auditory function. *Hearing Research*, *382*, 107779. https://doi.org/10.1016/j.heares.2019.107779, PubMed: 31505395

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*, 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x, PubMed: 20092565

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*, 349–359. https://doi.org/10.1152/jn.90896.2008, PubMed: 19439675

Maddox, R. K., & Lee, A. K. C. (2018). Auditory brainstem responses to continuous natural speech in human listeners. *eNeuro*, *5*, ENEURO.0441-17.2018. https://doi.org/10.1523/ENEURO.0441-17.2018, PubMed: 29435487

Madsen, S. M. K., Whiteford, K. L., & Oxenham, A. J. (2017). Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. *Scientific Reports*, *7*, 12624. https://doi.org/10.1038/s41598-017-12937-9, PubMed: 28974705

Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research*, *266*, 36–51. https://doi.org/10.1016/j.heares.2009.09.012, PubMed: 19788920

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *367*, 919–931. https://doi.org/10.1098/rstb.2011.0355, PubMed: 22371614

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., et al. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *Journal of Neuroscience*, *29*, 15564–15574. https://doi.org/10.1523/JNEUROSCI.3065-09.2009, PubMed: 20007480

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*, 1697–1706. https://doi.org/10.1093/cercor/bht355, PubMed: 24429136

Oxenham, A. J. (2008). Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants. *Trends in Amplification*, *12*, 316–331. https://doi.org/10.1177/1084713808325881, PubMed: 18974203

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*, e1001251. https://doi.org/10.1371/journal.pbio.1001251, PubMed: 22303281

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*, 1497–1503. https://doi.org/10.1111/j.1460-9568.2012.08060.x, PubMed: 22462504

Pressnitzer, D., Meddis, R., Delahaye, R., & Winter, I. M. (2001). Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus. *Journal of Neuroscience*, *21*, 6377–6386. https://doi.org/10.1523/JNEUROSCI.21-16-06377.2001, PubMed: 11487661

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology*, *18*, 1124–1128. https://doi.org/10.1016/j.cub.2008.06.053, PubMed: 18656355

Ross, B., Tremblay, K. L., & Alain, C. (2020). Simultaneous EEG and MEG recordings reveal vocal pitch elicited cortical gamma oscillations in young and older adults. *Neuroimage*, *204*, 116253. https://doi.org/10.1016/j.neuroimage.2019.116253, PubMed: 31600592

Skoe, E., & Kraus, N. (2010). Auditory brainstem reponse to complex sounds: A tutorial. *Ear and Hearing*, *31*, 302–324. https://doi.org/10.1097/AUD.0b013e3181cdb272, PubMed: 20084007

Sohmer, H., Pratt, H., & Kinarti, R. (1977). Sources of frequency following responses (FFR) in man. *Electroencephalography and Clinical Neurophysiology*, *42*, 656–664. https://doi.org/10.1016/0013-4694(77)90282-6, PubMed: 67025

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*, 155–166. https://doi.org/10.1162/jocn_a_01467, PubMed: 31479349

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data—A practical approach. *Journal of Neuroscience Methods*, *250*, 34–46. https://doi.org/10.1016/j.jneumeth.2014.08.002, PubMed: 25128257

Winkler, I., Takegata, R., & Sussman, E. (2005). Event-related brain potentials reveal multiple stages in the perceptual organization of sound. *Cognitive Brain Research*, *25*, 291–299. https://doi.org/10.1016/j.cogbrainres.2005.06.005, PubMed: 16005616

Winkler, I., Teder-Sälejärvi, W. A., Horváth, J., Näätänen, R., & Sussman, E. (2003). Human auditory cortex tracks task-irrelevant sound sources. *NeuroReport*, *14*, 75–118. https://doi.org/10.1097/00001756-200311140-00009, PubMed: 14600496

Wong, D. D. E., Fuglsang, S. A., Hjortkjaer, J., Ceolini, E., Slaney, M., & de Cheveigné, A. (2018). A comparison of regularization methods in forward and backward models for auditory attention decoding. *Frontiers in Neuroscience*, *12*, 531. https://doi.org/10.3389/FNINS.2018.00531, PubMed: 30131670

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience*, *32*, 855–869. https://doi.org/10.1080/23273798.2016.1262051

Yamagishi, S., Otsuka, S., Furukawa, S., & Kashino, M. (2016). Subcortical correlates of auditory perceptual organization in humans. *Hearing Research*, *339*, 104–111. https://doi.org/10.1016/j.heares.2016.06.016, PubMed: 27371867