Hearing Research 361 (2018) 92–102

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Low- and high-frequency cortical brain oscillations reflect dissociable mechanisms of concurrent speech segregation in noise

Anusha Yellamsetty [a], Gavin M. Bidelman [a, b, c, *]

[a] *School of Communication Sciences & Disorders, University of Memphis, Memphis, TN, USA*
[b] *Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA*
[c] *Univeristy of Tennessee Health Sciences Center, Department of Anatomy and Neurobiology, Memphis, TN, USA*

ARTICLE INFO

Article history:
Received 24 May 2017
Received in revised form
9 December 2017
Accepted 12 January 2018
Available online 2 February 2018

Keywords:
EEG
Time-frequency analysis
Double-vowel segregation
F0-benefit
Speech-in-noise perception

ABSTRACT

Parsing simultaneous speech requires listeners use pitch-guided segregation which can be affected by the signal-to-noise ratio (SNR) in the auditory scene. The interaction of these two cues may occur at multiple levels within the cortex. The aims of the current study were to assess the correspondence between oscillatory brain rhythms and determine how listeners exploit pitch and SNR cues to successfully segregate concurrent speech. We recorded electrical brain activity while participants heard double-vowel stimuli whose fundamental frequencies (F0s) differed by zero or four semitones (STs) presented in either clean or noise-degraded (+5 dB SNR) conditions. We found that behavioral identification was more accurate for vowel mixtures with larger pitch separations but F0 benefit interacted with noise. Time-frequency analysis decomposed the EEG into different spectrotemporal frequency bands. Low-frequency (θ, β) responses were elevated when speech did not contain pitch cues (0ST > 4ST) or was noisy, suggesting a correlate of increased listening effort and/or memory demands. Contrastively, γ power increments were observed for changes in both pitch (0ST > 4ST) and SNR (clean > noise), suggesting high-frequency bands carry information related to acoustic features and the quality of speech representations. Brain-behavior associations corroborated these effects; modulations in low-frequency rhythms predicted the speed of listeners' perceptual decisions with higher bands predicting identification accuracy. Results are consistent with the notion that neural oscillations reflect both automatic (pre-perceptual) and controlled (post-perceptual) mechanisms of speech processing that are largely divisible into high- and low-frequency bands of human brain rhythms.

1. Introduction

In normal auditory scenes (e.g., cocktail parties), listeners must parse acoustic mixtures to extract the intended message of a target, a process known as source segregation. Previous studies have suggested that fundamental frequency (F0) (i.e., pitch) differences provide a robust cue for identifying the constituents of concurrent speech. For instance, using synthetic double-vowel stimuli in a concurrent speech identification task, studies have shown that accuracy of identifying both vowels improves with increasing pitch differences between the vowels for F0 separations from 0 to about 4 semitones (STs) (Assmann and Summerfield, 1989; Assmann and Summerfield, 1990; Assmann and Summerfield, 1994; de Cheveigné et al., 1997). This improvement has been referred to as the "F0 benefit" (Arehart et al., 1997; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016). Thus, psychophysical research from the past several decades confirms that human listeners exploit F0 (pitch) differences to segregate concurrent speech.

Neural responses to concurrent speech and non-speech sounds have been measured at various levels of the auditory system including single-unit recordings in animals (Palmer, 1990; Portfors and Sinex, 2005; Sinex et al., 2003; Snyder and Sinex, 2002) and in human, via evoked potentials (Alain et al., 2005; Bidelman, 2017; Bidelman and Alain, 2015b; Dyson and Alain, 2004) and fMRI (Arnott et al., 2005). The segregation of complex signals is thought to involve a multistage hierarchy of processing, whereby initial pre-attentive processes partition the sound waveform into distinct acoustic features (e.g., pitch, harmonicity) which is then acted upon by later, post-perceptual Gestalt principles (Koffka, 1935) [e.g., grouping by physical similarity, temporal proximity, good continuity (Bregman, 1990)] and phonetic template matching (Alain

* Corresponding author. School of Communication Sciences & Disorders, University of Memphis, 4055 North Park Loop, Memphis, TN 38152, USA.
E-mail address: g.bidelman@memphis.edu (G.M. Bidelman).

https://doi.org/10.1016/j.heares.2018.01.006
0378-5955/© 2018 Elsevier B.V. All rights reserved.

et al., 2005; Meddis and Hewitt, 1992).

In humans, the neural correlates of concurrent speech segregation have been most readily studied using event-related brain potentials (ERPs). Modulations in ERP amplitude/latency provide an index of the timing and level of processing for emergent mechanisms of speech segregation. Mapping the time course of concurrent speech processing, modulations in neural activity have been observed as early as ~150–200 ms, indicative of pre-attentive signal detection, with conscious identification of simultaneous speech occurring slightly later, ~350–400 ms post-stimulus onset (Alain et al., 2007, 2005, 2017; Bidelman and Yellamsetty, 2017; Du et al., 2010; Reinke et al., 2003). Further perceptual learning studies have shown enhancements in the ERPs with successful learning in double vowel tasks in the form of an earlier and larger N1-P2 complex (enhanced sensory coding < 200 ms) coupled with larger slow wave activity (~400 ms), indicative of more effective cognitive processing/memory template matching (Alain et al., 2007; Reinke et al., 2003). Using brain-imaging methods (PET, fMRI), the spatial patterns of neural activation associated with speech processing have also been visualized in various regions of the auditory cortex (Giraud et al., 2004; Pulvermüller, 1999). For example, fMRI implicates a left thalamocortical network including thalamus, bilateral superior temporal gyrus and left anterior temporal lobe in successful double-vowel segregation (Alain et al., 2005).

One of the main factors affecting the parsing of simultaneous speech is signal-to-noise ratio (SNR). In real-world listening environments, successful recognition of noise-degraded speech is thought to reflect a frontotemporal speech network involving a close interplay between primary auditory sensory areas and inferior frontal brain regions (Bidelman and Alain, 2015b; Bidelman and Howell, 2016; Binder et al., 2004; Eisner et al., 2010). Consequently, dynamic F0 cues and noise SNR are likely to interact during the extraction of multiple auditory streams and occur relatively early (within few hundred milliseconds) in the neural hierarchy (Bidelman, 2017; Bidelman and Yellamsetty, 2017).

While prior studies have shed light on cortical activity underlying the neural encoding of concurrent speech, they cannot speak to how different frequency bands of the EEG (i.e., neural oscillations) relate to concurrent speech segregation. These frequency-specific “brain rhythms” become apparent only after averaging single-trial epochs in the spectral domain. The resulting neural spectrogram can be decomposed into various frequency bands which are thought to reflect local (high-frequency) and long-range (low -frequency) communication between different neural populations. Studies also suggest that various frequency ranges of the EEG may reflect different mechanisms of processing, including attention (Lakatos et al., 2008), navigation (Buzsáki, 2005), memory (Palva et al., 2010; Sauseng et al., 2008), motor planning (Donoghue et al., 1998), and speech-language comprehension (Doelling et al., 2014; Ghitza, 2011, 2013; Ghitza et al., 2013; Haarmann et al., 2002; Shahin et al., 2009). Although still debated, the general consensus is that lower frequency oscillations are associated with the perception, cognition, and action, whereas high-frequency bands are associated with stimulus transduction, encoding, and feature selection (von Stein and Sarnthein, 2000).

With regard to speech listening, different oscillatory activity may contribute to the neural coding of acoustic features in the speech signal or different internal cognitive operations related to the perceptual segregation process. Speech can be decomposed into different bands of time-varying modulations (i.e., slow-varying envelope vs. fast-varying fine structure) which are captured in the neural phase-locked activity of the scalp EEG (Bidelman, 2016). Theoretical accounts of brain organization suggest that different time-varying units of the speech signal (e.g., envelope vs. fine structure; phoneme vs. sentential segments) might be “tagged” by different frequency ranges of neural oscillations that coordinate brain activity at multiple spatial and temporal scales across distant cortical regions. Of relevance to speech coding, delta band (<3 Hz) oscillations have been shown to reflect processing related to sequencing syllables and words embedded within phrases (Ghitza, 2011, 2012). Theta (θ: 4–8 Hz) band has been linked with syllable coding at the word level (Bastiaansen et al., 2005; Giraud and Poeppel, 2012; Goswami, 2011) and attention/arousal (Aftanas et al., 2001; Paus et al., 1997). In contrast, beta (β: 15–30 Hz) band has been associated with the extraction of global phonetic features (Bidelman, 2015a, 2017; Fujioka et al., 2012; Ghitza, 2011), template matching (Bidelman, 2015a), lexical semantic memory access (Shahin et al., 2009), and perceptual binding in brain networks (Aissani et al., 2014; Brovelli et al., 2004; von Stein and Sarnthein, 2000). Lastly, gamma (γ: > 50 Hz) band has been associated with detailed phonetic features (Goswami, 2011), short duration cues (Giraud and Poeppel, 2012; Zhou et al., 2016), local network synchronization (Giraud and Poeppel, 2012; Haenschel et al., 2000), perceptual object construction (Tallon-Baudry and Bertrand, 1999), and experience-dependent enhancements in speech processing (Bidelman, 2017). Yet, the role of rhythmic neural oscillations in concurrent speech perception and how various frequency bands of the EEG relate to successful auditory scene analysis remains unclear.

In the present study, we aimed to further elucidate the neural mechanisms of concurrent speech segregation from the perspective of *oscillatory* brain activity. To this end, we recorded neuroelectric responses as listeners performed a double-vowel identification task during stimulus manipulations designed to promote or deny successful segregation (i.e., changes in F0 separation of vowels; with/without noise masking). Time-frequency analysis of the EEG provided novel insight into the correspondence between brain rhythms and speech perception and how listeners exploit pitch and SNR cues for successful segregation. Based on previous investigations on evoked (ERP) correlates of concurrent speech segregation (Alain et al., 2007; Bidelman and Yellamsetty, 2017; Reinke et al., 2003) we expected early modulations in higher frequency bands of the EEG (e.g., γ-band) would be sensitive to changes in F0-pitch and the SNR of speech. This would be consistent with the hypothesis that high frequency oscillations tag information related to the acoustic features of stimuli and the quality of speech representations. Additionally, we hypothesized that lower bands of oscillation (e.g., θ-band) would reflect more domain general, internal operations related to the perceptual segregation process and task demands (e.g., attention, listening effort, memory demands).

2. Methods

2.1. Subjects

Thirteen young adults (mean ± SD age: 26.1 ± 3.8 years; 10 females, 3 males) participated in the experiment. All had obtained a similar level of formal education (19.6 ± 2.8 years), were right handed (>43.2 laterality) (Oldfield, 1971), had normal hearing thresholds (i.e., ≤25 dB HL) at octave frequencies between 250 and 8000 Hz, and reported no history of neuropsychiatric disorders. Each gave written informed consent in compliance with a protocol approved by the University of Memphis Institutional Review Board.

2.2. General speech-in-noise recognition task

We measured listeners' speech-in-noise (SIN) recognition using the standardized QuickSIN test (Killion et al., 2004). We have previously shown a strong correspondence between QuickSIN scores and speech ERPs (Bidelman and Howell, 2016), justifying the inclusion of this instrument. Participants heard two lists embedded in four-talker babble noise, each containing six sentences with five key words. Sentences were presented at 70 dB SPL using pre-recorded signal-to-noise ratios (SNRs) which decreased in 5 dB steps from 25 dB (easy) to 0 dB (difficult). After each presentation, participants repeated the sentence and the number of correct key words were scored. “SNR loss” (computed in dB) was determined by subtracting the total number of correctly recalled words from 25.5. This metric represents the SNR required to correctly identify 50% of the key words across the sentences (Killion et al., 2004). SNR loss was measured for two lists separately for the left and right ear. The four responses were then averaged to obtain a stable SIN recognition score for each participant.

2.3. Electrophysiological procedures

2.3.1. Double vowel stimuli

Speech stimuli were modeled after previous studies on concurrent double-vowel segregation (Alain et al., 2007; Assmann and Summerfield, 1989; Assmann and Summerfield, 1990; Bidelman and Yellamsetty, 2017). Synthetic, steady-state vowel tokens (/a/ ,/i/, and/u/) were created using a Klatt synthesizer (Klatt, 1980) implemented in MATLAB® 2015b (The MathWorks, Inc.). Each token was 200 ms in duration including 10-ms $\cos^2$ onset/offset ramping. Vowel F0 and formant frequencies were held constant over the duration. F0 was either 100 or 125 Hz. Double-vowel stimuli were then created by combining single-vowel pairs. Each vowel pair had either identical (0 ST) or different F0s (4ST). That is, one vowel's F0 was set at 100 Hz while the other vowel had an F0 of 100 or 125 Hz so as to produce double-vowels with an F0 separation of either 0 or 4 semitones (STs). Each vowel was paired with every other vowel (except itself), resulting in a total of 6 unique double-vowel pairings (3 pairs x 2 F0 combinations). Double-vowels were presented in a clean and noise condition (separate blocks), in which stimuli were delivered concurrently with a backdrop of multi-talker noise babble (+5 dB SNR) (Bidelman and Howell, 2016; Nilsson et al., 1994). SNR was manipulated by changing the level of the masker rather than the signal to ensure that SNR was not positively correlated with overall sound level (Bidelman and Howell, 2016; Binder et al., 2004). Babble was presented continuously to avoid time-locking it with the stimulus presentation. We chose continuous babble over other forms of acoustic interference (e.g., white noise) because it more closely mimics real-world listening situations and tends to have a larger effect on the auditory ERPs (Kozou et al., 2005).

Stimulus presentation was controlled by MATLAB routed to a TDT RP2 interface (Tucker-Davis Technologies). Speech stimuli were delivered binaurally at a level of 81 dB SPL through ER-2 insert earphones (Etymotic Research). During EEG recording, listeners heard 50 exemplars of each double-vowel combination and were asked to identity *both* vowels as quickly and accurately as possible on the keyboard. Feedback was not provided. The inter-stimulus interval was jittered randomly between 800 and 1000 ms (20-ms steps, rectangular distribution) to avoid rhythmic entrainment of the EEG (Luck, 2005, p. 168) and listeners anticipating subsequent trials. The next trial commenced following the listener's behavioral response. The order of vowel pairs was randomized within and across participants; clean and noise conditions were run in separate blocks. A total of six blocks (3 clean, 3 noise) were completed, yielding 150 trials for each of the individual double-vowel conditions. Listeners were given 2–3 min breaks after each block (10–15 min after 3 blocks) as needed to avoid fatigue.

Prior to the experiment proper, we required that participants be able to identify single vowels in a practice run with >90% accuracy (e.g., Alain et al., 2007). This ensured their task performance would be mediated by *concurrent* sound segregation skills rather than isolated identification, *per se.*

2.3.2. EEG data recording and preprocessing

EEG recording procedures followed well-established protocols in our laboratory (Bidelman, 2015b; Bidelman and Howell, 2016; Bidelman and Yellamsetty, 2017). Neuroelectric activity was recorded from 64 sintered Ag/AgCl electrodes at standard 10-10 locations around the scalp (Oostenveld and Praamstra, 2001). Contact impedances were maintained <5 kΩ throughout the duration of the experiment. EEGs were digitized using a sampling rate of 500 Hz (SynAmps RT amplifiers; Compumedics Neuroscan). Electrodes placed on the outer canthi of the eyes and the superior and inferior orbit were used to monitor ocular activity. The data were pre-processed by thresholding EEG amplitudes at ±100 μV. Ocular artifacts (saccades and blink artifacts) were then corrected in the continuous EEG using a principal component analysis (PCA) (Wallstrom et al., 2004). Data were visually inspected for bad channels and paroxysmal electrodes were interpolated from the adjacent four nearest neighbor channels (distance weighted). These procedures helped remove myogenic and other artifacts prior to time-frequency analysis that can affect the interpretation of oscillatory responses (Pope et al., 2009). During online acquisition, all electrodes were referenced to an additional sensor placed ~1 cm posterior to Cz. Data were re-referenced off-line to a common average reference. EEGs were then epoched (-200-1000 ms), baseline-corrected to the pre-stimulus interval, and digitally filtered (1–100 Hz, zero-phase) for response visualization and time-frequency analysis. To obtain an adequate number of trials for analysis, we pooled responses to collapse across different vowel pairs. This yielded 450 trials per listener for the four conditions of interest [i.e., 2 SNRs (clean, noise) x 2 F0s (0 ST, 4 ST)]. The entire experimental protocol including behavioral and electrophysiological testing took ~2 h to complete.

2.4. EEG time-frequency analysis

Evoked potential (ERP) results related to this dataset are reported in our companion paper (Bidelman and Yellamsetty, 2017). New time-frequency analyses (applied here) were used to evaluate the correspondence between *rhythmic* brain oscillations and speech perception and how listeners exploit pitch and SNR cues for successful segregation.

From epoched EEGs, we computed time-frequency decompositions of single-trial data to assess frequency-specific changes in oscillatory neural power (Bidelman, 2015a, 2017). For each trial epoch, the time-frequency map (i.e., spectrogram) was extracted using Mortlet wavelets as implemented in the MATLAB package Brainstorm (Tadel et al., 2011). This resulted in an estimate of the power for each time-frequency point over the bandwidth (1–100 Hz; 1 Hz steps) and time course (-200–1000 ms) of each epoch window. Using the Mortlet basis function, spectral resolution decreased linearly with increasing frequency; the full width half maximum (FWHM) was ~1 Hz near DC and approached ~20 Hz at 60 Hz. Temporal resolution improved exponentially with increasing frequency; FWHM was ~3 s near DC and ~50 ms at 60 Hz. Single-trial spectrograms were then averaged across trials to obtain time-frequency maps for each subject and stimulus condition (see Fig. 2). When power is expressed relative to the baseline pre-stimulus interval (-200–0 ms), these spectrographic maps are known as event-related spectral perturbations (ERSPs) (Delorme and Makeig, 2004). ERSPs represent the increase/decrease in EEG spectral power relative to the baseline pre-stimulus period (in dB). They contain neural activity that is both time- and phase-locked to the eliciting stimulus (i.e., evoked activity) as well as non-phase-locked responses (i.e., induced oscillatory activity) generated by the ongoing stimulus presentation (Bidelman, 2015a, 2017; Shahin et al., 2009; Trainor et al., 2009). To reduce the dimensionality of the data, we restricted our analysis to the Fz electrode. This channel is ideal for measuring auditory evoked responses (Picton et al., 1999b) and time-frequency oscillations (Bidelman, 2015a, 2017) to speech which are both maximal over frontocentral scalp locations. Moreover, scalp topographies of our data (pooled across subjects and conditions) confirmed that most band responses were strongest near frontocentral regions of the scalp (see Fig. 3). While we restrict subsequent analyses to Fz, it should be noted that in pilot testing, we also analyzed responses at different electrode clusters. However, results were qualitatively similar to those reported herein (data not shown).

To quantify frequency-specific changes in oscillatory power to concurrent speech, we extracted time courses from ERSP maps in five different bands. Band-specific waveforms were extracted by taking "slices" of the ERSP maps averaged across different frequency ranges: 5–7 Hz (θ), 8–12 Hz (α), 15–29 Hz (β), 30–59 Hz (γ_{low}), and 60–90 Hz (γ_{high}). This resulted in a running time waveform within each prominent frequency band of the EEG, similar to an ERP. We then contrasted band-specific waveforms (i.e., clean vs. noise; 0 ST vs. 4 ST) to compare the neural encoding of double-vowel stimuli across the main factors of interest (i.e., SNR and pitch). We used a running permutation test (EEGLAB's statcond function; Delorme and Makeig, 2004) to determine the time points over which band activity differed between stimulus conditions ($p < 0.05$, $N = 1000$ resamples). We required that segments persisted contiguously for ≥25 ms to be considered reliable and help control false positives (Chung and Bidelman, 2016; Guthrie and Buchwald, 1991).

This initial analysis revealed time segments where band-specific oscillations were modulated by our stimulus manipulations (i.e., SNR and pitch). To better quantify stimulus-related changes, we extracted peak power from the mid-point of the time segments showing significant differences in band activity: θ: 450 ms; β: 350 ms; $\gamma_{low/high}$: average of peak power at 25 and 175 ms (see Fig. 3). Grand average ERSP scalp topographies (pooled across stimulus conditions) are shown for each band in Fig. 3. Scalp maps confirmed that synchronized responses to speech mixtures were maximal over the frontocentral plane (Alain et al., 2006; Picton et al., 1999a).

2.5. Behavioral data analysis

2.5.1. Identification accuracy and the "F0 benefit"

Behavioral identification was analyzed as the percent of trials where *both* vowel sounds were identified correctly. For statistical analyses, %-correct scores were arcsine transformed to improve homogeneity of variance assumptions necessary for parametric statistics (Studebaker, 1985). Increasing the F0 between two vowels provides a pitch cue which leads to an improvement in accuracy identifying concurrent vowels (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992)—an effect referred to as the "F0-benefit" (Arehart et al., 1997; Bidelman and Yellamsetty, 2017; Chintanpalli and Heinz, 2013). To provide a singular measure of double-vowel identification we calculated the F0-benefit for each listener, computed as the difference in performance (%-correct) between the 4ST and 0ST conditions. F0-benefit was computed separately for clean and noise stimuli allowing us to compare the magnitude of F0 benefit in concurrent speech segregation with and without noise interference.

2.5.2. Reaction time (RTs)

Behavioral speech labeling speeds [i.e., reaction times (RTs)], were computed separately for each participant as the median response latency across trials for a given double-vowel condition. RTs were taken as the time lapse between the onset of the stimulus presentation and listeners' identification of both vowel sounds. Following our previous studies on the neural correlates of speech perception (e.g., Bidelman and Walker, 2017; Bidelman et al., 2013), RTs shorter than 250 ms or exceeding 6000 ms were discarded as implausibly fast responses and lapses of attention, respectively.

2.6. Statistical analysis

Unless otherwise noted, two-way, mixed-model ANOVAs were conducted on all dependent variables (GLIMMIX Procedure, SAS® 9.4, SAS Institute, Inc.). Stimulus SNR (2 levels; clean, +5 dB noise) and semitones (2 levels; 0ST, 4ST) functioned as fixed effects; subjects served as a random factor. Tukey-Kramer multiple comparisons controlled Type I error inflation. An *a priori* significance level was set at $\alpha = 0.05$.

To examine the degree to which neural responses predicted behavioral speech segregation, we performed weighted least square regression between listeners' band-specific amplitudes and (i) their accuracy, and RTs in the double-vowel task and (ii) QuickSIN scores. Robust bisquare fitting was achieved using "fitlm" in MATLAB. To arrive at a comparable and single measure to describe how neurophysiological responses distinguished speech using pitch cues, we derived a "*neural* F0 benefit," computed as the difference between each listener's 4ST and 0ST responses. As in behavioral F0 benefit, this neural analogue was computed separately for the clean and noise conditions. We then regressed behavioral and neural F0 benefits to assess brain-behavior correspondences. We reasoned that listeners who experience larger changes in their neural encoding of speech with added pitch cues (i.e., stronger neural F0 benefit) would have larger behavioral gains in the double-vowel segregation from 0 to 4 ST (i.e., experience bigger perceptual F0 benefit).

3. Results

3.1. Behavioral data

Behavioral speech identification accuracy and RTs for double-vowel segregation are shown in Fig. 1 A. Listeners obtained near-ceiling performance (96.9 ± 1.4%) when identifying single vowels. In contrast, double-vowel identification was considerably more challenging; listeners' accuracy ranged from ~30 to 70% depending on the presence of noise and pitch cues. An ANOVA conducted on behavioral accuracy confirmed a significant SNR x F0 interaction [$F_{1, 12} = 5.78$, $p = 0.0332$], indicating that successful double-vowel identification depended on both the noise level and presence of F0 pitch cues. Post hoc contrasts revealed listeners showed a similar level of performance with and without noise for 0 ST vowels, those which did not contain pitch cues. Performance increased ~30% across the board with greater F0 separations (i.e., 4ST > 0ST). F0-benefit was larger for clean relative to +5 dB SNR speech [$t_{12} = 2.15$, $p = 0.026$ (one-tailed)], suggesting listeners made stronger use of pitch cues when segregating clean compared to acoustically impoverished speech.

Analysis of RTs revealed a marginal effect of SNR [$F_{1, 12} = 4.11$, $p = 0.065$]; listeners tended to be slower identifying clean compared to noisy speech (Fig. 1B). The slowing of RTs coupled with better %-identification for clean compared to noise-degraded speech is indicative of a time-accuracy tradeoff in concurrent sound segregation. Indeed, RTs and %-correct scores were highly correlated [$r = 0.46$, $p = 0.006$] such that more accurate identification corresponded with slower decisions.

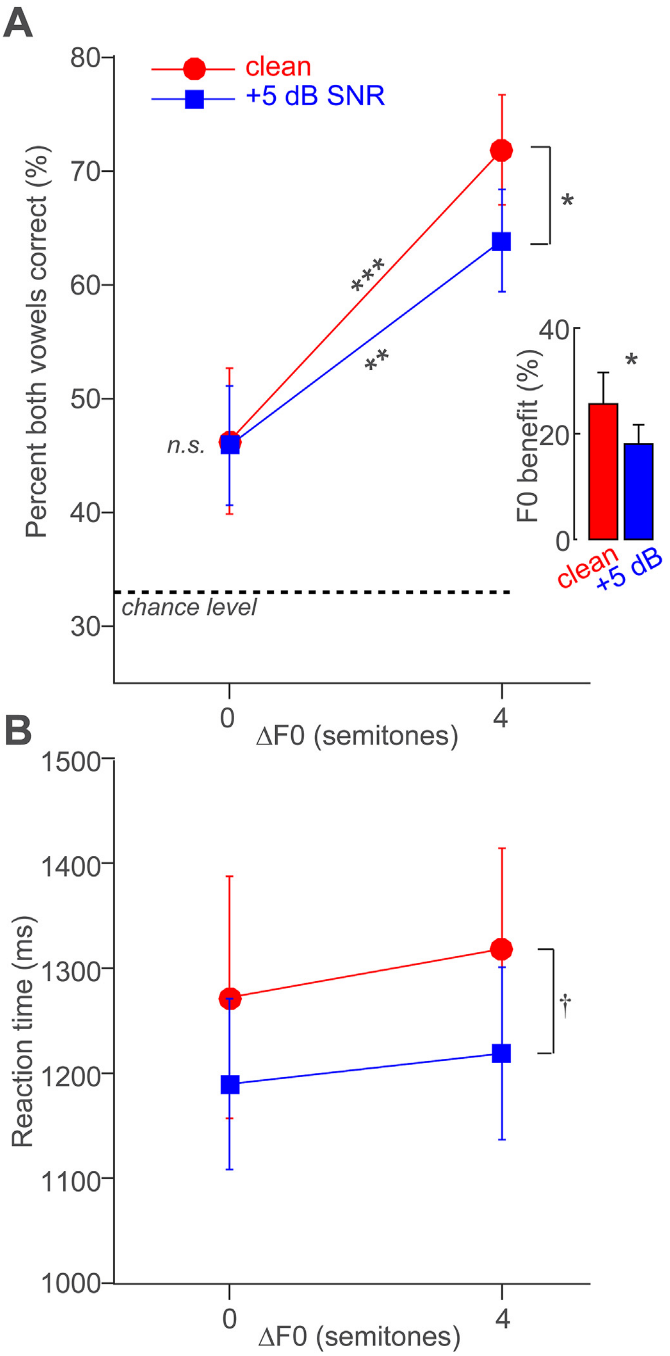

Fig. 1. Behavioral responses for segregating double-vowel stimuli. (A) Accuracy for identifying both tokens of a two-vowel mixture. Performance is poorer when concurrent speech sounds contain the same F0 (0ST) and improve ~30% when vowels contain differing F0s (4ST). (*Insert*) Behavioral F0-benefit, defined as the improvement in %-accuracy from 0ST to 4ST, indexes the added benefit of pitch cues to speech segregation. F0-benefit is stronger for clean vs. noisy (+5 dB SNR) speech indicating that listeners are poorer at exploiting pitch cues when segregating acoustically-degraded signals. (B) Speed (i.e., RTs) for double-vowel segregation. Listeners are marginally faster at identifying speech in noise. However, faster RTs at the expense of poorer accuracy (panel A) suggests a time-accuracy tradeoff in double-vowel identification. Data reproduced from Bidelman and Yellamsetty (2017). error bars = ±1 s. e.m.

3.2. Neural oscillatory responses during double-vowel coding

Grand average ERSP time-frequency maps are shown for each of the noise and ST conditions in Fig. 2. Fig. 3 shows time waveforms for the 5–7 Hz (θ), 8–12 Hz (α), 15–29 Hz (β), 30–59 Hz (γ_{low}), and 60–90 Hz (γ_{high}) bands extracted from the spectrographic maps of Fig. 2. Each reflects how different frequency oscillations in the EEG code double-vowel mixtures. Generally speaking, lower frequency bands including θ- and α-band showed sustained activity over the duration of the trial which appeared stronger for more difficult stimulus conditions (i.e., noisy speech and 0ST conditions). Compared to clean speech, β-band activity also appeared larger (more positive) ~400–500 ms after speech onset. Lastly, higher γ-band showed broadband transient activations that seem to tag the onset (see 25 ms) and offset (see 200 ms) of the evoking speech stimulus (cf. Ross et al., 2010). These high γ-band events also appeared stronger for clean relative to noise-degraded speech and for 0ST vs. 4ST vowel mixtures. In terms of the overall time course of spectral responses, the strong modulations of high γ-band in clean and at 0ST were followed by negative modulation of β-band and sustained positive modulation of the θ-band. The directions of these band amplitude effects were reversed in the noise and 4 ST conditions.

Fig. 3C shows the SNR × ST interaction waveforms. Interactions were confined to α- and β-bands, at early (~150–200 ms) time windows after stimulus onset. These early interactions replicate (are consistent with) the noise × pitch interactions observed in the N1-P2 time window of our previous ERP study on double-vowel coding (Bidelman and Yellamsetty, 2017) and thus, were not explored further.

Next, we aimed to quantify changes in spectral band power due to each acoustic factor (SNR, STs). For each band time course for the two main effects (i.e., Fig. 3 and B), peak amplitudes were extracted from the temporal center of segments showing significant stimulus-related modulations based on initial permutation tests (see ■, Fig. 3A–B). For θ-band (Fig. 4A), we found elevated spectral responses when speech did not contain pitch cues (i.e., 0ST > 4ST) [$F_{1,\ 36} = 0.413$, $p = 0.0495$], whereas the β-band and γ_{low} -band (Fig. 4B and C), showed stronger oscillatory activity for clean speech (i.e., clean > noise) [β band: $F_{1,\ 36} = 9.73$, $p = 0.0036$; γ_{low} band: $F_{1,\ 36} = 5.15$, $p = 0.0294$]. Modulations in γ_{high} power oscillations (Fig. 4D) were observed for changes in both pitch (0ST > 4ST) [$F_{1,\ 36} = 5.65$, $p = 0.0229$] and SNR (clean > noise) [$F_{1,\ 36} = 16.87$; $p = 0.0002$]. Together, these findings demonstrate that difference in neural activity to speech between conditions is derived by acoustic features, signal quality, and the cognitive effort which causes changes in underlying low vs. high bands of oscillatory activity.

3.3. Brain-behavior relationships

Bivariate regressions between band-specific EEG amplitudes and behavioral accuracy and RTs are shown in Fig. 5A and B, respectively. For each frequency band, we derived a singular measure of *neural* F0-benefit, computed as the change in response with and without pitch cues (e.g., $\Delta\ \beta_{4ST\ -\ }\beta_{0ST}$). This neural measure was then regressed against each listener's *behavioral* F0-benefit for the accuracy and RT measures (i.e., $\Delta\ PC_{4ST\ -\ }PC_{0ST}$ for accuracy scores; $\Delta\ RT_{4ST\ -\ }RT_{0ST}$ for reaction times). Paralleling our previous work on speech processing (cf. Bidelman, 2017; Bidelman and Walker, 2017), we reasoned that larger neural differentiation between the 0ST and 4ST would correspond to larger gains in behavioral performance (i.e., larger perceptual F0-benefit). Repeating this analysis for each band allowed us to evaluate potential mechanistic differences in how different neural rhythms map to behavior. Each matrix cell shows the regression's *t*-statistic which indicates both the magnitude and sign (i.e., negative vs. positive) of the association between variables.

These analyses revealed that γ_{low} was associated ($R^2 = 0.17$) with %-accuracy in the double vowel task when pooling clean and noise conditions. Analysis by SNR indicated that this correspondence was driven by how γ_{low} differentiated clean speech ($R^2 = 0.42$). Additional links were found between behavioral RT

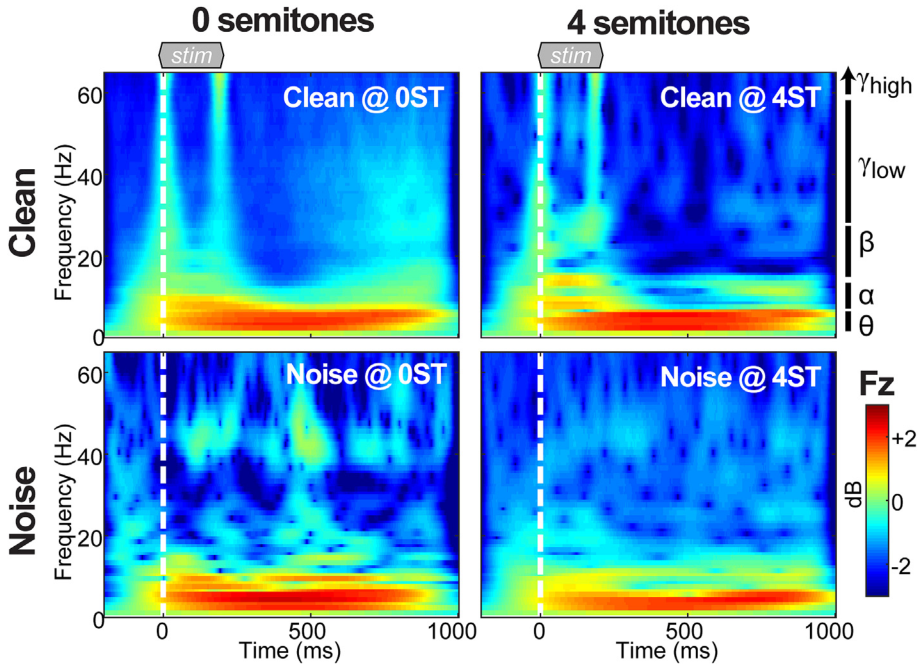

Fig. 2. Neural oscillatory responses to concurrent speech sounds are modulated by SNR and the presence/absence of pitch cues. ERSP time-frequency maps (Fz channel) quantify both "evoked" and "rhythmic" changes in EEG power relative to the baseline period. Each panel represents the response to double-vowel stimuli with (4ST) or without (0ST) a difference in voice fundamental frequency for stimuli presented either in clean or +5 dB SNR of noise. Light gray regions above the spectrograms show the schematized stimulus. Dotted lines, stimulus onset ($t = 0$).

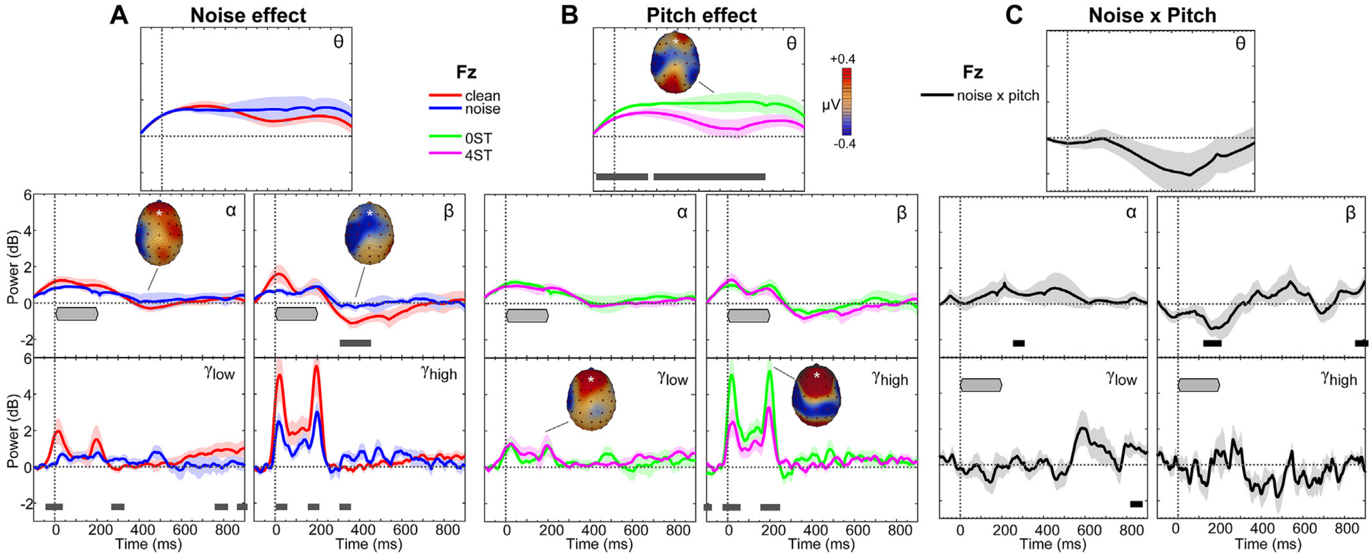

Fig. 3. Band-specific time courses during double-vowel segregation. Shown here are response time courses for each frequency band of the EEG extracted from ERSP spectrograms and their interaction. Band waveforms contrast how noise SNR (**A**), F0 pitch (**B**), and their interaction (**C**; SNR x pitch) affect the neural encoding of double-vowel mixtures. A permutation test shows contiguous segments (≥25 ms duration) where spectral power differs between stimulus conditions (■ segments; $p < 0.05$, $N = 1000$ resamples). Modulations in β- and high γ-band distinguish clean from noise-degraded speech (β: clean < noise; γ_{high} = clean > noise). Contrastively, pitch cues are distinguished by modulations in the θ band (0ST > 4ST) and γ_{high} band (0ST > 4ST). Head maps (pooled across stimulus conditions and subjects) show the topographic distribution of each band across the scalp at time points where the band-specific effects are largest. * Fz electrode for subsequent analysis. Gray regions, schematized stimulus. Shading = ±1 s. e.m.

speeds and neural F0-benefit, particularly for low-frequency bands of the EEG. Notably, changes in θ- ($R^2 = 0.71$) and β- ($R^2 = 0.19$) oscillations predicted listeners' RTs, particularly for noise-degraded speech.[1] Collectively, these findings imply that higher frequency oscillatory rhythms (γ-band) might reflect the quality of stimulus representation and thus accuracy in identifying double-vowel mixtures. In contrast, low-frequency oscillations are associated with the speed of individuals' decisions and thus the listening effort associated with concurrent sound processing.

Listeners QuickSIN scores were low (−0.73 ± 1.3 dB SNR loss), consistent with the average speech-in-noise perception abilities for normal-hearing listeners (i.e., 0 dB). QuickSIN scores were not

[1] We carried out the same correlations using average amplitude across the entirety of significant band segments (see ■, Fig. 3). These brain-behavior correlations were qualitatively similar to the peak analysis results shown in Fig. 5.

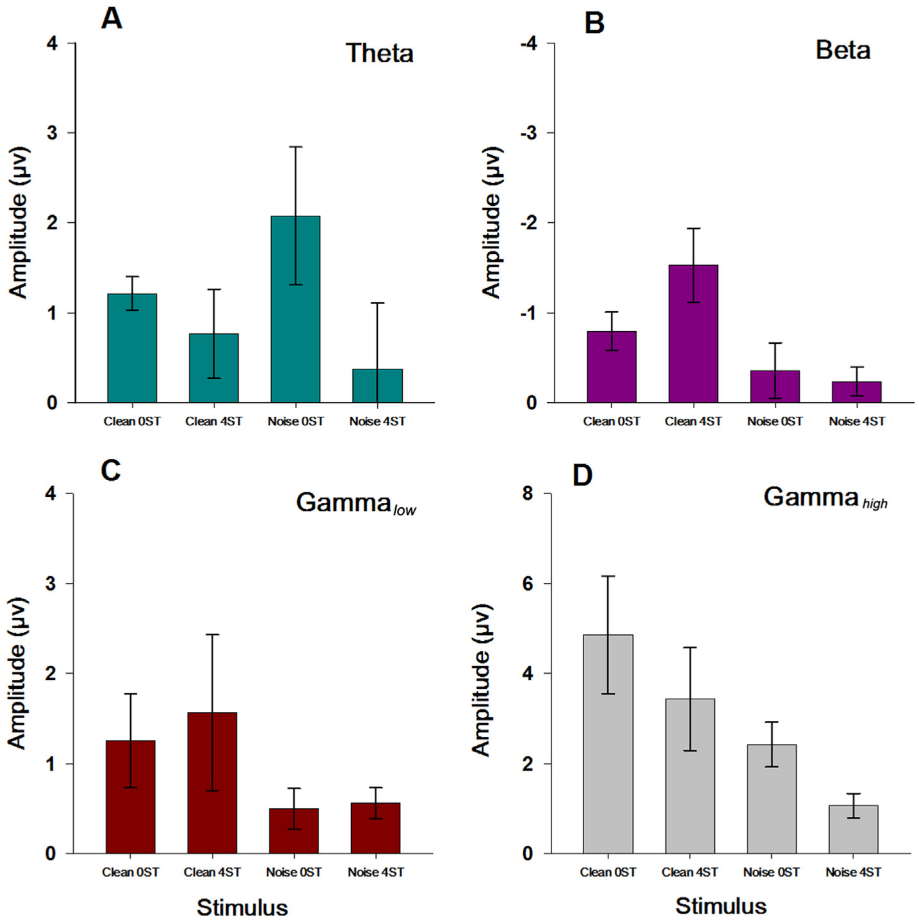

Fig. 4. Band-specific mean spectral peak amplitudes across conditions. Shown here are mean amplitudes for each frequency band extracted from the temporal center of segments showing significant stimulus-related modulations (see Fig. 3). (A) θ-band spectral responses were elevated when speech did not contain pitch cues (i.e., 0ST > 4ST). (B) β-band and (C) γ_{low} -band showed stronger desynchronization for clean compared to noise-degraded speech (i.e., clean > noise). Note that negative is plotted up for this band. (D) γ_{high} power modulations were observed for changes in both pitch (0ST > 4ST) and SNR (clean > noise). error bars = ±1 s. e.m.

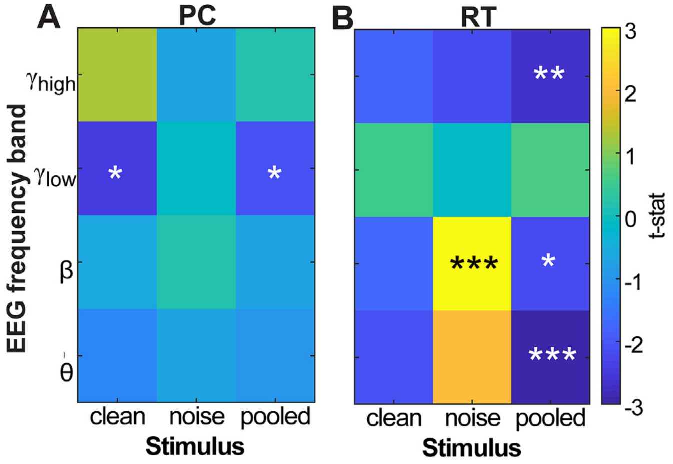

Fig. 5. Brain-behavior correlations underlying double-vowel segregation. Individual cells of each matrix show the *t*-statistic for the regression indicating both the magnitude and sign of association between *neural* F0-benefit and listeners' corresponding *behavioral* F0-benefit. In both cases, larger F0-benefit reflects more successful neural/behavioral speech segregation with the addition of pitch cues (i.e., 4ST > 0ST). (A) Correspondences between neural responses and identification accuracy (%); (B) correspondence with RTs. Changes in γ_{low} activity predict improved behavioral accuracy in double-vowel identification whereas the speed of listeners' decision are predicted by changes in lower oscillations (θ and β band). PC = percent correct, RT = reaction times. $^{*}p < 0.05$, $^{**}p \le 0.01$, $^{***}p \le 0.001$.

correlated with any band-specific oscillations across SNR or pitch conditions.

4. Discussion

The present study measured rhythmic neuroelectric brain activity as listeners rapidly identified double-vowel stimuli varying in their voice pitch (F0) and noise level (SNR). Results showed three primary findings: (i) behaviorally, listeners exploit F0 differences between vowels to segregate speech and this perceptual F0 benefit is larger for clean compared to noise degraded (+5 dB SNR) stimuli; (ii) oscillatory power of lower θ and β frequency bands of the EEG reflects cognitive processing modulated by task demands (e.g., listening effort, memory), whereas high γ_{low} and γ_{high} -band power tracks acoustic features (e.g., envelope) and quality (i.e., noisiness) of the speech signal that reflect stimulus encoding; (iii) perceptual performance in segregating speech sounds is predicted by modulatory effects in different bands: low-frequency oscillations correlate with behavioral reaction times in double vowel identification whereas high-frequency oscillations are linked to accuracy. The differential changes in power across frequency bands of the EEG suggest the engagement of different brain mechanisms supporting speech segregation that vary with pitch and noise cues in auditory mixtures.

4.1. Effects of SNR and F0 cues on behavioral concurrent vowel segregation

Consistent with previous behavioral data (Arehart et al., 1997; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016; Reinke et al., 2003), we found that listeners were better at perceptually identifying speech mixtures when vowels contained different F0s (4ST) compared to identical (0ST) F0s in both clean and noise conditions (clean > noise). This perceptual F0 benefit was larger for clean compared to noise degraded (+5 dB SNR) stimuli. However, we extend prior studies by demonstrating that the acoustic stressor of noise limits the effectiveness of these pitch cues for segregation. Indeed, F0-benefit was weaker for double-vowel identification amidst noise compared to clean listening conditions. Similarly, smaller ΔRTs (accompanied by lower accuracy) for segregating in noise suggests that listeners experienced a time-accuracy tradeoff such that they achieved more accurate identification of speech at the expense of slower decision times (Fig. 1).

Computationally, the identification of concurrent vowels is thought to involve a two-stage process in which the auditory system first determines the number of elements present in a mixture (i.e., "1" vs. "2" sounds) and then seeks their identities (~150–200 ms). The former process (segregation) is thought to involve a comparison of the incoming periodicities of double-vowel F0s, which could be realized via autocorrelation-like mechanisms in peripheral (Bidelman and Alain, 2015a; Chintanpalli et al., 2014; Du et al., 2010; Meddis and Hewitt, 1992) and/or auditory cortical neurons (Alain et al., 2005; Bidelman and Alain, 2015a; Du et al., 2010).

Indeed, neurons in primary and surrounding belt areas of auditory cortex are both sensitive to pitch and even display multi-peaked tuning with peaks occurring at harmonically-related frequencies (Bendor et al., 2012; Kikuchi et al., 2014). Following F0-based segregation, the process of determining vowel identity could be realized via template matching mechanisms (~300–400 ms) in which each representation is matched against internalized memory profiles for both vowel constituents. Using a computational model of this two-stage model (i.e., autocorrelation-based segregation followed by template matching), Meddis and colleagues (Meddis and Hewitt, 1992) have shown that identification of two vowels improves from ~40% to 70% when they differ in F0 from 0 to 4 ST—consistent with the F0-benefit in this study. While F0 cues are likely the primary cue for segregation in our double vowel task, conceivably, listeners might also use additional acoustic cues to parse speech such as spectral differences associated with formants (Chintanpalli and Heinz, 2013), temporal envelope cues produced by harmonic interactions (Culling and Darwin, 1993), and spectral edges.

4.2. Cortical oscillations reveal mechanistic differences in concurrent speech segregation divisible by frequency band

It is useful to cast our behavioral data in the context of this computational framework. We found that listeners showed weaker F0-benefit when speech was presented in noise. Poor performance in the noise conditions could result either from poorer segregation at the initial front end (prior to classification) or weaker matching between the noisy vowel representations and their speech templates. Our behavioral data do not allow us to unambiguously adjudicate these two explanations. In this regard, EEG time-frequency results help isolate different mechanistic accounts. In response to a stimulus, synchronous temporal activity is represented as multiple time courses in brain networks via EEG oscillations whose amplitude depends on the degree of neural synchrony. Different frequencies respond differently to sensory stimuli and task demands (Hanslmayr et al., 2011). Stimulus rhythmic event-related activity can either increase (synchronization) or decrease (de-synchronization) as networks are either engaged or disengaged, respectively (Destexhe et al., 2007).

Presumably, the acoustic features contributing to the segregation of the speech depend on the availability of those cues to the auditory system. That is, the encoding and weighting of acoustic cues along the auditory pathway may change depending on the quality of the incoming signal. Electrophysiologically, we observed multiple, frequency-specific time courses to concurrent speech segregation with activity unfolding within different spectral channels of the EEG dependent on both the pitch and SNR of speech. Previous M/EEG work has shown similar sequences of events in the early object negativity response (~150 ms) (Alain et al., 2005; Du et al., 2010) and early interactions of pitch and noise cues (~200 ms) (Bidelman and Yellamsetty, 2017) followed by automatic registration of F0 differences at ~250 ms (Alain et al., 2005; Du et al., 2010).

In cases where vowel mixtures were further distorted by noise, γ_{high} power showed reduced tracking of stimulus onset/offset (cf. Ross et al., 2010). γ_{high} power was also stronger for 0ST compared to 4ST speech (i.e., mixtures which did not contain pitch cues). Higher γ activity for both clean and 0ST conditions may be due to the fact that these stimuli offer a more veridical and robust representation of the speech signal envelope; clean speech being unconcluded and 0ST vowels offering a singular harmonic structure (common F0). Under this interpretation, modulations in γ activity in our double vowel task are arguably ambiguous as they signal both cleaner signals (clean > noise) simultaneously with representations that cannot be cleanly segregated (0ST > 4ST) (cf. Fig. 3A and B). Relatedly, brain-behavior correlations showed that larger changes in γ activity with the addition of pitch cues were associated with *poorer* behavioral F0-benefit (Fig. 5A). Given that higher bands of oscillations are thought to reflect signal identity and the construction of perceptual objects (Bidelman, 2015a, 2017; Tallon-Baudry and Bertrand, 1999), our data suggest that the auditory brain must rely on more than object-based information for successful segregation.

In contrast to the higher γ-band modulations, we also observed distinct modulation in lower bands of the EEG that covaried with successful speech segregation. Interestingly, β band amplitudes were suppressed for easier stimulus conditions (e.g., clean 4ST; Fig. 4B), suggesting a desynchronization in this frequency range. Similarly, θ-band activity showed prominent increases (synchronization) for difficult 0ST and noise-degraded speech. β band (15–30 Hz) has been linked with the extraction of global phonetic features (Bidelman, 2015a, 2017; Fujioka et al., 2012; Ghitza, 2011), template matching (Bidelman, 2015a), lexical semantic memory access (Shahin et al., 2009), and perceptual binding (Aissani et al., 2014; Brovelli et al., 2004; von Stein and Sarnthein, 2000). In contrast, θ-band may reflect and attention/arousal (Aftanas et al., 2001; Paus et al., 1997). Enhancements in θ-activity and suppression in β-modulations are known to correlate with the level of attention and memory load in a wide variety of tasks (Bashivan et al., 2014; Bastiaansen et al., 2005; Fries et al., 2001). Modulations of M/EEG amplitudes during the conscious identification of simultaneous speech occurs around ~350–400 ms post stimulus onset (Alain, 2007; Alain et al., 2005, 2017; Bidelman and Yellamsetty, 2017; Du et al., 2010; Reinke et al., 2003) relating to the time course of β- and θ-band oscillatory activity observed in this study.

Thus, we suggest perceptual success in parsing multiple speech streams is driven by the degree of cognitive processing (e.g., attentional deployment, listening effort) that is determined by the availability of acoustic features and signal quality. Cleaner, less distorted speech presumably allows more successful matches between speech acoustics and internalized speech templates which would aid identification. This notion is supported by the fact that larger changes in θ responses were associated with *smaller* ΔRTs whereas larger changes in β responses were associated with *larger* ΔRTs (Fig. 5B). Given that listeners required a longer time to accurately judge double-vowels (i.e., $\Delta RT_{clean} > \Delta RT_{noise}$ time-accuracy tradeoff; Fig. 1B), the most parsimonious interpretation of our neural results are that θ-band elevates due to increased listening effort or cognitive demands of the task (e.g., conditions without F0 cues) whereas β-band decreases, reflecting easier and/or more successful memory template matching (e.g., clean speech conditions).

4.3. On the additivity vs. interactions of cues for concurrent sound segregation

Notably, while EEG measures showed a correspondence with behavior for double vowel identification, we did not observe correlations between neural measures and QuickSIN scores. However, this might be expected given differences in task complexity and the fact that the former was recorded during electrophysiological testing while the latter was not. Nevertheless, these findings corroborate our previous studies and suggest that mechanisms that exploit sequential and concurrent auditory streaming are likely independent (or at least different) from the mechanisms recruited for complex speech in noise recognition (Alain et al., 2017; Hutka et al., 2013). For example, the QuickSIN may rely more on cognitive (rather than perceptual) processes, such as attention, working memory, and linguistic processing, while double-vowel identification tasks used in the present study are more perceptual-based. Future work is needed to explore the relationship (or lack thereof) between concurrent speech segregation and more generalized speech-in-noise recognition tests.

The differential changes in oscillatory θ-, β-, and γ power and F0 x SNR interaction in α- and β-bands illustrates potential differences in the brain mechanisms supporting speech segregation that are largely divisible into high- and low-frequency brain rhythms. The neural interaction of pitch and noise that are circumscribed to α- and β-bands and in the earliest time windows (~150–200 ms) is consistent with our previous ERP studies which revealed significant F0 x SNR interactions in concurrent vowel encoding and perception in the timeframe of the N1-P2 complex (Bidelman and Yellamsetty, 2017). Overall, we found that different acoustic factors (SNR vs. noise) influenced the neural encoding of speech dynamically with interaction effects early but additive effects occurring later in time. Our results are partially in agreement with the additive effects on concurrent vowel perception shown by Du et al. (2011), who suggested that listeners rely on a linear summation of cues to accumulate evidence during auditory scene analysis. Indeed, our data show that high- (γ) and low- (θ) frequency responses carry independent information on speech processing later in time (>300–400 ms). However, our results further reveal that acoustic cues (here SNR and F0) can interact earlier (~100–200 ms; Fig. 3C) to impact double vowel processing. Notably, Du et al. (2011) study investigated the effects of F0 and *spatial location* on concurrent vowel perception. Given that spatial and non-spatial (cf. F0) cues are largely processed via independent information channels of the brain (i.e., dorsal and ventral pathways) (Arnott et al., 2004), acoustic differences among sources might be expected to combine linearly as reported in that study (Du et al., 2011). In contrast, our behavioral and electrophysical results suggest acoustic cues that affect the inherent acoustic representation of speech signals (i.e., pitch and noise) can actually interact fairly early in the time course of speech segregation and are not processed in a strictly additive manner (Bidelman and Yellamsetty, 2017).

4.4. Directions for future work

Previous ERP studies have shown success in identifying concurrent vowels improves with training accompanied by decreased N1 and P2 latencies and enhanced P2 peak amplitude (Alain, 2007; Alain et al., 2007). In future extensions of this work, it would be interesting to examine how the weighting of neural activity changes across frequency bands with perceptual learning. For example, a testable hypothesis is that neural changes in lower frequency bands might accompany top-down automatization during successful learning. We would also predict that higher frequency bands would begin showing improved signal coding with task repetition and increased familiarity with the incoming signal. Another interesting study would be to investigate multiple competing streams and how attention might modulate concurrent speech segregation (Ding and Simon, 2012; Krumbholz et al., 2007). Future studies are needed to test the role of band-specific mechanisms of the EEG in relation to short-term speech sound training, learning, and attentional effects on concurrent speech segregation.

5. Conclusions

By measuring time-frequency changes in the EEG during double vowel identification, we found band-specific differences in oscillatory spectral responses which seem to represent unique mechanisms of speech perception. Over the 200 ms stimulus duration, early envelope tracking of the stimulus duration (onset/offset) was observed in higher frequency oscillations of the γ band. This was followed by stronger desynchronization (suppression) in the mid-frequency β oscillations around (~250–350 ms). Finally, differences in lower frequency θ oscillations were more pervasive and persisted across a larger extent of each trial (~400–500 ms after stimulus onset). We infer that early portions of time-frequency activity (higher-bands) likely reflect pre-perceptual encoding of acoustic features and follow the quality of the speech signal. This capture of stimulus properties is then followed by post-perceptual cognitive operations (reflected in low EEG bands) that involve the degree of listening effort and task demands. Tentatively, we posit that successful speech segregation is governed by more accurate perceptual object construction, auditory template matching, and deceased listening effort/attentional allocation, indexed by the γ-, β-, and θ-band modulations, respectively.

Acknowledgements

The authors thank Amy Fehrenbach for assistance in data collection and Kelsey Mankel for comments on previous versions of this manuscript.

References

Aftanas, L., Varlamov, A., Pavlov, S., Makhnev, V., Reva, N., 2001. Affective picture processing: event-related synchronization within individually defined human theta band is modulated by valence dimension. Neurosci. Lett. 303, 115–118.
Aissani, C., Martinerie, J., Yahia-Cherif, L., Paradis, A.-L., Lorenceau, J., 2014. Beta, but not gamma, band oscillations index visual form-motion integration. PLoS One 9, e95541.
Alain, C., 2007. Breaking the wave: effects of attention and learning on concurrent sound perception. Hear. Res. 229, 225–236.
Alain, C., Snyder, J.S., He, Y., Reinke, K.S., 2006. Changes in auditory cortex parallel rapid perceptual learning. Cerebr. Cortex 17, 1074–1084.
Alain, C., Snyder, J.S., He, Y., Reinke, K.S., 2007. Changes in auditory cortex parallel rapid perceptual learning. Cerebr. Cortex 17, 1074–1084.
Alain, C., Reinke, K., He, Y., Wang, C., Lobaugh, N., 2005. Hearing two things at once: neurophysiological indices of speech segregation and identification. J. Cognit. Neurosci. 17, 811–818.
Alain, C., Arsenault, J.S., Garami, L., Bidelman, G.M., Snyder, J.S., 2017. Neural correlates of speech segregation based on formant frequencies of adjacent vowels. Sci. Rep. 7, 1–11.

Arehart, K.H., King, C.A., McLean-Mudgett, K.S., 1997. Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss. J. Speech Lang. Hear. Res. 40, 1434–1444.

Arnott, S.R., Binns, M.A., Grady, C.L., Alain, C., 2004. Assessing the auditory dual-pathway model in humans. Neuroimage 22, 401–408.

Arnott, S.R., Grady, C.L., Hevenor, S.J., Graham, S., Alain, C., 2005. The functional organization of auditory working memory as revealed by fMRI. J. Cognit. Neurosci. 17, 819–831.

Assmann, P.F., Summerfield, Q., 1989. Modeling the perception of concurrent vowels: vowels with the same fundamental frequency. J. Acoust. Soc. Am. 85, 327–338.

Assmann, P.F., Summerfield, Q., 1990. Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J. Acoust. Soc. Am. 88, 680–697.

Assmann, P.F., Summerfield, Q., 1994. The contribution of waveform interactions to the perception of concurrent vowels. J. Acoust. Soc. Am. 95, 471–484.

Bashivan, P., Bidelman, G.M., Yeasin, M., 2014. Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity. Eur. J. Neurosci. 40, 3774–3784.

Bastiaansen, M.C., Van Der Linden, M., Ter Keurs, M., Dijkstra, T., Hagoort, P., 2005. Theta responses are involved in lexical—semantic retrieval during language processing. J. Cognit. Neurosci. 17, 530–541.

Bendor, D., Osmanski, M.S., Wang, X., 2012. Dual-pitch processing mechanisms in primate auditory cortex. J. Neurosci. 32, 16149–16161.

Bidelman, G.M., 2015a. Induced neural beta oscillations predict categorical speech perception abilities. Brain Lang. 141, 62–69.

Bidelman, G.M., 2015b. Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. Hear. Res. 323, 68–80.

Bidelman, G.M., 2016. Relative contribution of envelope and fine structure to the subcortical encoding of noise-degraded speech. J. Acoust. Soc. Am. 140, EL358–363.

Bidelman, G.M., 2017. Amplified induced neural oscillatory activity predicts musicians' benefits in categorical speech perception. Neuroscience 348, 107–113.

Bidelman, G.M., Alain, C., 2015a. Hierarchical neurocomputations underlying concurrent sound segregation: connecting periphery to percept. Neuropsychologia 68, 38–50.

Bidelman, G.M., Alain, C., 2015b. Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception. J. Neurosci. 35, 1240–1249.

Bidelman, G.M., Howell, M., 2016. Functional changes in inter- and intra-hemispheric auditory cortical processing underlying degraded speech perception. Neuroimage 124, 581–590.

Bidelman, G.M., Yellamsetty, A., 2017. Noise and pitch interact during the cortical segregation of concurrent speech. Hear. Res. 351, 34–44.

Bidelman, G.M., Walker, B., 2017. Attentional modulation and domain specificity underlying the neural organization of auditory categorical perception. Eur. J. Neurosci. 45, 690–699.

Bidelman, G.M., Moreno, S., Alain, C., 2013. Tracing the emergence of categorical speech perception in the human auditory system. Neuroimage 79, 201–212.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., Ward, B.D., 2004. Neural correlates of sensory and decision processes in auditory object identification. Nat. Neurosci. 7, 295–301.

Bregman, A., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press, Cambridge, MA.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., Bressler, S.L., 2004. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. Proc. Natl. Acad. Sci. U. S. A. 101, 9849–9854.

Buzsáki, G., 2005. Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. Hippocampus 15, 827–840.

Chintanpalli, A., Heinz, M.G., 2013. The use of confusion patterns to evaluate the neural basis for concurrent vowel identification. J. Acoust. Soc. Am. 134, 2988–3000.

Chintanpalli, A., Ahlstrom, J.B., Dubno, J.R., 2014. Computational model predictions of cues for concurrent vowel identification. J. Assoc. Res. Oto 15, 823–837.

Chintanpalli, A., Ahlstrom, J.B., Dubno, J.R., 2016. Effects of age and hearing loss on concurrent vowel identification. J. Acoust. Soc. Am. 140, 4142.

Chung, W.-L., Bidelman, G.M., 2016. Cortical encoding and neurophysiological tracking of intensity and pitch cues signaling English stress patterns in native and nonnative speakers. Brain Lang. 155–156, 49–57.

Culling, J.F., Darwin, C., 1993. Perceptual separation of simultaneous vowels: within and across-formant grouping by F 0. J. Acoust. Soc. Am. 93, 3454–3467.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., Aikawa, K., 1997. Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. J. Acoust. Soc. Am. 101, 2839–2847.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Meth. 134, 9–21.

Destexhe, A., Hughes, S.W., Rudolph, M., Crunelli, V., 2007. Are corticothalamic 'up'states fragments of wakefulness? Trends Neurosci. 30, 334–342.

Ding, N., Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. Unit. States Am. 109, 11854–11859.

Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta—theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85, 761–768.

Donoghue, J.P., Sanes, J.N., Hatsopoulos, N.G., Gaál, G., 1998. Neural discharge and local field potential oscillations in primate motor cortex during voluntary movements. J. Neurophysiol. 79, 159–173.

Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., Alain, C., 2010. Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. Cerebr. Cortex 21, 698–707.

Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., Alain, C., 2011. Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. Cerebr. Cortex 21, 698–707.

Dyson, B.J., Alain, C., 2004. Representation of concurrent acoustic objects in primary auditory cortex. J. Acoust. Soc. Am. 115, 280–288.

Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., Scott, S.K., 2010. Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. J. Neurosci. 30, 7179–7186.

Fries, P., Reynolds, J.H., Rorie, A.E., Desimone, R., 2001. Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291, 1560–1563.

Fujioka, T., Trainor, L.J., Large, E.W., Ross, B., 2012. Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. J. Neurosci. 32, 1791–1802.

Ghitza, O., 2011. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front. Psychol. 2, 130.

Ghitza, O., 2012. On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. 3, 238.

Ghitza, O., 2013. The theta-syllable: a unit of speech information defined by cortical function. Front. Psychol. 4, 138.

Ghitza, O., Giraud, A.-L., Poeppel, D., 2013. Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. Front. Hum. Neurosci. 6, 340.

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

Giraud, A., Kell, C., Thierfelder, C., Sterzer, P., Russ, M., Preibisch, C., Kleinschmidt, A., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cerebr. Cortex 14, 247–255.

Goswami, U., 2011. A temporal sampling framework for developmental dyslexia. Trends Cognit. Sci. 15, 3–10.

Guthrie, D., Buchwald, J.S., 1991. Significance testing of difference potentials. Psychophysiology 28, 240–244.

Haarmann, H.J., Cameron, K.A., Ruchkin, D.S., 2002. Neural synchronization mediates on-line sentence processing: EEG coherence evidence from filler-gap constructions. Psychophysiology 39, 820–825.

Haenschel, C., Baldeweg, T., Croft, R.J., Whittington, M., Gruzelier, J., 2000. Gamma and beta frequency oscillations in response to novel auditory stimuli: a comparison of human electroencephalogram (EEG) data with in vitro models. Proc. Natl. Acad. Sci. Unit. States Am. 97, 7645–7650.

Hanslmayr, S., Gross, J., Klimesch, W., Shapiro, K.L., 2011. The role of alpha oscillations in temporal attention. Brain Res. Rev. 67, 331–343.

Hutka, S., Alain, C., Binns, M., Bidelman, G.M., 2013. Age-related differences in the sequential organization of speech sounds. J. Acoust. Soc. Am. 133, 4177–4187.

Kikuchi, Y., Horwitz, B., Mishkin, M., Rauschecker, J.P., 2014. Processing of harmonics in the lateral belt of macaque auditory cortex. Front. Neurosci. 8, 1–13.

Killion, M.C., Niquette, P.A., Gudmundsen, G.I., Revit, L.J., Banerjee, S., 2004. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 116, 2395–2405.

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 67, 971–995.

Koffka, K., 1935. Principles of Gestalt Psychology, International Library of Psychology, Philosophy and Scientific Method. Harcourt Brace, New York.

Kozou, H., Kujala, T., Shtyrov, Y., Toppila, E., Starck, J., Alku, P., Naatanen, R., 2005. The effect of different noise types on the speech and non-speech elicited mismatch negativity. Hear. Res. 199, 31–39.

Krumbholz, K., Eickhoff, S.B., Fink, G.R., 2007. Feature-and object-based attentional modulation in the human auditory "where" pathway. J. Cognit. Neurosci. 19, 1721–1733.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. science 320, 110–113.

Luck, S., 2005. An Introduction to the Event-related Potential Technique. MIT Press, Cambridge, MA, USA.

Meddis, R., Hewitt, M.J., 1992. Modeling the identification of concurrent vowels with different fundamental frequencies. J. Acoust. Soc. Am. 91, 233–245.

Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. 95, 1085–1099.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.

Oostenveld, R., Praamstra, P., 2001. The five percent electrode system for high-resolution EEG and ERP measurements. Clin. Neurophysiol. 112, 713–719.

Palmer, A., 1990. The representation of the spectra and fundamental frequencies of steady-state single-and double-vowel sounds in the temporal discharge patterns of Guinea pig cochlear-nerve fibers. J. Acoust. Soc. Am. 88, 1412–1426.

Palva, J.M., Monto, S., Kulashekhar, S., Palva, S., 2010. Neuronal synchrony reveals working memory networks and predicts individual memory capacity. Proc. Natl. Acad. Sci. Unit. States Am. 107, 7580–7585.

Paus, T., Zatorre, R.J., Hofle, N., Caramanos, Z., Gotman, J., Petrides, M., Evans, A.C., 1997. Time-related changes in neural systems underlying attention and arousal during the performance of an auditory vigilance task. J. Cognit. Neurosci. 9, 392–408.

Picton, T., Alain, C., Woods, D., John, M., Scherg, M., Valdes-Sosa, P., Bosch-Bayard, J., Trujillo, N., 1999a. Intracerebral sources of human auditory-evoked potentials. Audiol. Neurotol. 4, 64–79.

Picton, T.W., Alain, C., Woods, D.L., John, M.S., Scherg, M., Valdes-Sosa, P., Bosch-Bayard, J., Trujillo, N.J., 1999b. Intracerebral sources of human auditory-evoked potentials. Audiol. Neuro. Otol. 4, 64–79.

Pope, K.J., Fitzgibbon, S.P., Lewis, T.W., Whitham, E.M., Willoughby, J.O., 2009. Relation of gamma oscillations in scalp recordings to muscular activity. Brain Topogr. 22, 13–17.

Portfors, C.V., Sinex, D.G., 2005. Coding of Communication Sounds in the Inferior Colliculus, the Inferior Colliculus. Springer, pp. 411–425.

Pulvermüller, F., 1999. Words in the brain's language. Behav. Brain Sci. 22, 253–279.

Reinke, K., He, Y., Wang, C., Alain, C., 2003. Perceptual learning modulates sensory evoked response during vowel segregation. Cognit. Brain Res. 17, 781–791.

Ross, B., Schneider, B., Snyder, J.S., Alain, C., 2010. Biological markers of auditory gap detection in young, middle-aged, and older adults. PLoS One 5, e10101.

Sauseng, P., Klimesch, W., Gruber, W.R., Birbaumer, N., 2008. Cross-frequency phase synchronization: a brain mechanism of memory matching and attention. Neuroimage 40, 308–317.

Shahin, A.J., Picton, T.W., Miller, L.M., 2009. Brain oscillations during semantic evaluation of speech. Brain Cognit. 70, 259–266.

Sinex, D.G., Guzik, H., Li, H., Sabes, J.H., 2003. Responses of auditory nerve fibers to harmonic and mistuned complex tones. Hear. Res. 182, 130–139.

Snyder, R.L., Sinex, D.G., 2002. Immediate changes in tuning of inferior colliculus neurons following acute lesions of cat spiral ganglion. J. Neurophysiol. 87, 434–452.

Studebaker, G.A., 1985. A "rationalized" arcsine transform. J. Speech Lang. Hear. Res. 28, 455–462.

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M., 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci. 2011, 1–13.

Tallon-Baudry, C., Bertrand, O., 1999. Oscillatory gamma activity in humans and its role in object representation. Trends Cognit. Sci. 3, 151–162.

Trainor, L.J., Shahin, A.J., Roberts, L.E., 2009. Understanding the benefits of musical training: effects on oscillatory brain activity. Ann. N. Y. Acad. Sci. 1169, 133–142.

von Stein, A., Sarnthein, J., 2000. Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization. Int. J. Psychophysiol. 38, 301–313.

Wallstrom, G.L., Kass, R.E., Miller, A., Cohn, J.F., Fox, N.A., 2004. Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. Int. J. Psychophysiol. 53, 105–119.

Zhou, H., Melloni, L., Poeppel, D., Ding, N., 2016. Interpretations of frequency domain analyses of neural entrainment: periodicity, fundamental frequency, and harmonics. Front. Hum. Neurosci. 10.