Hearing Research 351 (2017) 34–44

Contents lists available at ScienceDirect

Hearing Research

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

Research Paper

Noise and pitch interact during the cortical segregation of concurrent speech

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

[a] School of Communication Sciences & Disorders, University of Memphis, Memphis, TN, 38152, USA
[b] Institute for Intelligent Systems, University of Memphis, Memphis, TN, 38152, USA
[c] Univeristy of Tennessee Health Sciences Center, Department of Anatomy and Neurobiology, Memphis, TN, 38163, USA

ARTICLE INFO

Article history:
Received 24 March 2017
Received in revised form
9 May 2017
Accepted 23 May 2017
Available online 25 May 2017

Keywords:
EEG/ERPs
Double-vowel segregation
F0-benefit
Speech-in-noise perception
Neural classification

ABSTRACT

Behavioral studies reveal listeners exploit intrinsic differences in voice fundamental frequency (F0) to segregate concurrent speech sounds—the so-called “F0-benefit.” More favorable signal-to-noise ratio (SNR) in the environment, an extrinsic acoustic factor, similarly benefits the parsing of simultaneous speech. Here, we examined the neurobiological substrates of these two cues in the perceptual segregation of concurrent speech mixtures. We recorded event-related brain potentials (ERPs) while listeners performed a speeded double-vowel identification task. Listeners heard two concurrent vowels whose F0 differed by zero or four semitones presented in either clean (no noise) or noise-degraded (+5 dB SNR) conditions. Behaviorally, listeners were more accurate in correctly identifying both vowels for larger F0 separations but F0-benefit was more pronounced at more favorable SNRs (i.e., pitch × SNR interaction). Analysis of the ERPs revealed that only the P2 wave (~200 ms) showed a similar F0 x SNR interaction as behavior and was correlated with listeners' perceptual F0-benefit. Neural classifiers applied to the ERPs further suggested that speech sounds are segregated neurally within 200 ms based on SNR whereas segregation based on pitch occurs later in time (400−700 ms). The earlier timing of extrinsic SNR compared to intrinsic F0-based segregation implies that the cortical extraction of speech from noise is more efficient than differentiating speech based on pitch cues alone, which may recruit additional cortical processes. Findings indicate that noise and pitch differences interact relatively early in cerebral cortex and that the brain arrives at the identities of concurrent speech mixtures as early as ~200 ms.

1. Introduction

To properly analyze the auditory scene and endure the “cocktail party,” listeners must exploit various acoustic cues to segregate concurrent sounds. The process of auditory streaming is thought to rely on several acoustic principles including (among other factors) the degree of (in)harmonicity (Alain et al., 2001; Bidelman and Alain, 2015a), temporal coherence/(a)synchrony (Van Noorden, 1975), spectral content, and spatial configuration between multiple auditory objects (for reviews, see Bidet-Caulet and Bertrand, 2009; Bregman, 1990; Oxenham, 2008; Shamma et al., 2011). In particular, differences in the fundamental frequency (F0) between two or more sounds (i.e., pitch cues) represents one of the more robust acoustic factors for perceptual segregation. Auditory stimuli containing the same F0 are perceived as a single perceptual object whereas multiple F0s tend to promote hearing multiple sources. F0-based segregation is thought to reflect the grouping of spectral components that originate from a common target signal and the fact that F0 changes allow the auditory system to track time-varying properties of the voice over time (Assmann, 1996).

To probe the perceptual segregation of concurrent *speech* mixtures, behavioral studies have generally employed synthetic double-vowel stimuli (Assmann and Summerfield, 1989; Assmann and Summerfield, 1990; de Cheveigné et al., 1997a; de Cheveigné

* 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).

http://dx.doi.org/10.1016/j.heares.2017.05.008
0378-5955/© 2017 Elsevier B.V. All rights reserved.

et al., 1997b; Parsons, 1976). In these paradigms, listeners hear two simultaneous vowels and are asked to correctly identify both tokens. Findings of these studies show that speech identification accuracy improves with increasing pitch differences between vowels for F0 separations from 0 to about 4 semitones (STs)—an improvement referred to as the "F0-benefit" (Arehart et al., 1997; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016). From previous psychophysical studies, it is clear that listeners exploit intrinsic F0 differences to successfully segregate competing speech. Unfortunately, the neurobiological mechanisms underlying the separation and subsequent identification of overlapping speech remains poorly understood.

Segregation of speech and non-speech signals is thought to reflect a complex, distributed neural network involving both subcortical and cortical brain regions (Alain et al., 2005b; Bidelman and Alain, 2015a; Dyson and Alain, 2004; Palmer, 1990; Sinex et al., 2002). In humans, functional magnetic resonance imaging (fMRI) implicates a left thalamo-cortical network including thalamus, bilateral superior temporal gyrus, and left anterior temporal lobe in successful double-vowel segregation (Alain et al., 2005b). Event-related brain potentials (ERPs) have further delineated the time course of concurrent speech processing, with modulations in neural activity ~150–200 ms and 350–400 ms after sound onset (Alain et al., 2005a, 2007; Reinke et al., 2003)[1]. The sensitivity of neural responses in both an early and late time frame supports the notion of a multistage model of concurrent speech processing in which the spectral signatures of each vowel are extracted automatically in early auditory sensory cortex (or even subcotically; Meddis and Hewitt, 1992; Palmer, 1990) and then matched against their respective phonetic templates in memory shortly thereafter (Alain et al., 2005a; Bregman, 1990). This proposition is further bolstered by perceptual learning studies, which show that more successful learners in double-vowel tasks show enhancements in their ERPs in the form of decreased latencies and increased amplitudes of the N1-P2 complex (enhanced sensory coding) and larger slow wave activity around ~400 ms (more efficient cognitive processing/memory template matching) (Alain et al., 2007, 2015; Reinke et al., 2003).

Another important issue in understanding concurrent speech segregation is the effect of extrinsic acoustic interferences (e.g., external noise). Real-world listening environments (e.g., classrooms, cocktail parties, restaurants) nearly always contain some degree of background interference (Helfer and Wilber, 1990) and listeners must parse noise from target signals to achieve robust understanding. Additive noise tends to obscure less intense portions of the speech signal, reduce its signal-to-noise ratio (SNR), and prevent audible access to salient speech cues normally exploited for comprehension (e.g., temporal envelope; Bidelman, 2016; Shannon et al., 1995; Swaminathan and Heinz, 2012). In terms of the neural substrates of degraded speech processing, ERP studies demonstrate that noise weakens and prolongs the cortical encoding of (isolated) speech sounds dependent on signal SNR (Bidelman and Howell, 2016; Billings et al., 2009, 2010). Degraded speech perception is thought to reflect a fronto-temporal speech network involving a close interplay between primary auditory sensory areas and inferior frontal brain regions (Bidelman and Dexter, 2015; Bidelman and Howell, 2016; Binder et al., 2004; Du et al., 2014; Eisner et al., 2010).

Conceivably, pitch-based cues and noise could interact during the extraction of multiple auditory streams. For instance, tracking dynamic F0 cues may help listeners decide which fluctuations belong to target speech vs. interfering signals (Qin and Oxenham, 2005). This could aid the monitoring of auditory sources (Assmann, 1996) and improve speech perception in background noise (Bidelman and Krishnan, 2010; Macdonald et al., 2010; Nabelek et al., 1989). Despite the importance of F0 cues (Alain et al., 2007; Assmann and Summerfield, 1990; Chintanpalli et al., 2014; de Cheveigné et al., 1997b; Meddis and Hewitt, 1992; Parsons, 1976) and SNR (Bidelman and Howell, 2016; Billings et al., 2013) to successful speech perception, we are unaware of any studies examining how composite noise and pitch information affect the parsing of simultaneous speech.

To elucidate the neural mechanisms and time course of concurrent speech segregation, we recorded neuroelectric brain responses as listeners performed a double-vowel identification task. Consistent with previous ERP studies using similar paradigms (Alain et al., 2007; Reinke et al., 2003), we hypothesized that F0 cues would be carried via early auditory cortical activity within the first ~250 ms after the initiation of speech. However, we extend previous studies on concurrent speech processing by characterizing the effects of additive acoustic noise on double-vowel segregation. Given the importance of low-frequency, F0-based cues to noise-degraded speech perception (Bidelman and Krishnan, 2010; Macdonald et al., 2010; Nabelek et al., 1989), we hypothesized that noise (SNR) and pitch (F0) cues might interact during the segregation of speech mixtures, producing a differential effect on its neural encoding. To provide a novel, more fine-grained analysis of the temporal emergence of speech segregation, we also applied multivariate classification techniques to classify neural responses and identify the earliest time at which brain activity differentiated speech based on intrinsic (F0-cues) and extrinsic (SNR) acoustic properties of speech mixtures.

2. Methods

2.1. Participants

Thirteen young adults (mean ± SD age: 26.1 ± 3.8 years; 10 female, 3 male) 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 were native speakers of American English. None reported a history of neuropsychiatric disorders. On average, listeners had 5.6 ± 3.6 years of formal musical training. However, this is well below the criteria of most music-related plasticity studies which generally define "musicians" as individuals with a decade or more of experience (Bidelman and Krishnan, 2010; Bidelman et al., 2014; Parbery-Clark et al., 2009a; Zendel and Alain, 2012). Each participant gave informed written consent in compliance with a protocol approved by the University of Memphis Institutional Review Board.

Generalized speech-in-noise (SIN) recognition skills were assessed using the QuickSIN (Killion et al., 2004). We have previously shown strong correspondence between QuickSIN scores and neurophysiological responses (Bidelman and Bhagat, 2015; Bidelman and Howell, 2016). Participants heard sentence lists embedded in four-talker babble noise. "SNR loss" was computed from the number of keywords recalled and represents the SNR required to correctly identify 50% of the target items (Killion et al., 2004). QuickSIN scores were measured for four lists and averaged

[1] Though our focus is on *concurrent vowel identification* paradigms (Alain et al., 2005a; Reinke et al., 2003), it should be noted that the time course of concurrent sound segregation depends on the stimulus paradigm. Less perceptually taxing sounds including nonspeech harmonic tone complexes (Alain et al., 2001; Bidelman and Alain, 2015a) and competing streams of talkers (Ding and Simon, 2012) can elicit neural correlates of sound segregation that occur within ~100 ms.

to obtain a stable estimate of each listener's real-world SIN recognition skills.

2.2. Electrophysiological procedures

2.2.1. 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). Synthetic, steady-state vowel tokens (/a/, /i/, and /u/) were created using a Klatt synthesizer (Klatt, 1980) implemented in MATLAB® 2015b (The MathWorks, Inc., Natick, MA). Each token was 200 ms in duration including 10-ms $\cos^2$ onset/offset ramping to prevent spectral splatter. F0 and formant frequencies were held constant over the vowel duration. The F0 was either 100 or 125 Hz. Double-vowel stimuli were created by combining pairs of single-vowels where each pair had either identical (0 semitones; STs) 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. Each vowel was paired with every other vowel (except itself), resulting in a total of 6 unique double-vowel stimuli (i.e., 3 pairs x 2 F0 combinations). Double-vowels were presented in both a clean and noise condition (separate blocks) in which stimuli were delivered concurrent with a backdrop of multi-talker noise babble (+5 dB SNR) (Bidelman and Howell, 2016; Nilsson et al., 1994). This noise level was selected based on extensive pilot testing which showed this SNR hindered double vowel identification but avoided floor performance. Masker level was manipulated rather than the signal to ensure that SNR was not positively correlated with overall sound level (Bidelman and Howell, 2016; Binder et al., 2004). The babble was presented continuously to avoid time-locking with the stimulus presentation. We chose babble over other forms of acoustic inference (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, Alachua, FL). Speech stimuli were delivered binaurally at an intensity of 81 dB SPL through ER-2 insert earphones (Etymotic Research, Elk Grove, IL). During each block of 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 and listeners were told ahead of time that every trial would contain two unique vowels. The interstimulus 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. Order of vowel pairs was randomized within and across participants and 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. The entire experimental protocol including behavioral and electrophysiological testing took ~2 h to complete.

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.*

2.2.2. ERP recording and preprocessing

EEG recording procedures followed well-established protocols from our laboratory (Bidelman, 2015; Bidelman and Chung, 2015; Bidelman and Howell, 2016). Neuroelectric activity was recorded from 64 sintered Ag/AgCl electrodes at standard 10-10 scalp locations (Oostenveld and Praamstra, 2001). Contact impedances were maintained <5 kΩ. EEGs were digitized using a sampling rate of 500 Hz (SynAmps RT amplifiers; Compumedics Neuroscan, Charlotte, NC). Electrodes placed on the outer canthi of the eyes and the superior and inferior orbit were used to monitor ocular activity. Saccade and blink artifacts were then corrected in the continuous EEG using principal component analysis (PCA) (Wallstrom et al., 2004). 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 (CAR). EEGs were then epoched (-200-1000 ms), baseline-corrected to the pre-stimulus interval, and digitally filtered (1–30 Hz, zero-phase) for response visualization and ERP analysis. To obtain an adequate number of trials for ERP analysis, we pooled responses to collapse across the different vowel pairs. This yielded 450 trials per listener for each of the four conditions of interest [i.e., 2 SNRs (clean, noise) x 2 F0s (0 ST, 4 ST)].

2.2.3. Topographic ANOVA (TANOVA)

In initial exploratory analysis, we used a topographic ANOVA (TANOVA) to identify the spatial and temporal points at which ERPs were sensitive to our stimulus manipulations (i.e., SNR and STs) (for details, see Koenig and Melie-Garcia, 2010; Murray et al., 2008). TANOVAs were implemented in the Curry 7 Neuroimaging Suite (Compumedics Neuroscan). The TANOVA used a randomization procedure ($N = 1000$ resamples) that tested the distribution of the ERP's topography in the measured data against a surrogate distribution, derived by exchanging all participants and electrodes in the data. The percentage of shuffled cases where the effect size obtained after randomization was equal to or larger than the measured effect size obtained in the observed data provided an estimate of the probability of the null hypothesis. This analysis yielded running *p*-values across the epoch that identified the time points at which ERPs were significantly modulated by the main (SNR, ST) and interaction effects (SNR x ST) of our stimuli. We used the *within* option in Curry 7 which implements the randomization tests within subjects and the comparisons are conducted across subjects, akin to a repeated measures ANOVA with paired comparisons.

2.2.4. ERP peak quantification

For the purpose of data reduction and to minimize potential bias in electrode selection, we collapsed a subset of the 64-channel sensor data into a single region of interest (ROI) encompassing a cluster of six frontocentral electrodes (Fp1, Fpz, Fp2, F1, Fz, F2). This ROI was guided by our previous reports on the neural correlates of speech perception, which found that speech ERPs were most prominent at frontocentral scalp locations, indicative of bilateral sources in the Sylvian fissure (Bidelman and Lee, 2015; Bidelman and Walker, 2017; Bidelman et al., 2013, 2014).

Peak amplitudes and latencies were measured for the prominent deflections (N1, P2) within this ROI cluster. Visual inspection of the responses indicated that P1 was weak and could not be reliably measured at the single-subject level. Following conventions in previous studies (Bidelman and Howell, 2016; Irimajiri et al., 2005), N1 was taken as the minimum negativity between 85 and 150 ms and P2 as maximum positivity between

150 and 260 ms. In addition, the initial TANOVA and visual inspection of grand average traces revealed a late modulation with changes in ST that peaked around ~400 ms (see Fig. 5). This late wave (LW) was quantified as the peak positivity between the 350–450 ms.

2.2.5. *Discriminant function analysis: predicting concurrent speech segregation via ERP classification*

As a finer grained measure of speech processing, we built several neural classifiers to identify the *spatiotemporally* emergence of speech segregation. We have previously used similar machine learning and multivariate techniques to "decode" perceptual, pathological, and stimulus-related events in the EEG (e.g., Bidelman et al., 2017; Lee and Bidelman, in press). For each stimulus contrast of interest (i.e., pitch or SNR-based segregation), we built a time-varying neural classifier based on a sliding window analysis (15 ms steps) over the ERP's time course. Within each window, a linear discriminant analysis (LDA) function was used to segregate each listener's neural responses into mutually exclusive groups based on amplitude differences at that sample. For example, the SNR classifier attempted to predict a given response (ERP_X) as being evoked by either a "clean" or "noise" stimulus. Classification accuracy was then assessed by determining the proportion of responses accurately predicted against the actual eliciting stimulus (ground truth). Repeating this procedure across time provided a running accuracy of the classifier's performance. This approach was repeated for each electrode location and listener to examine the *spatiotemporal* discrimination of speech using solely ERP amplitude differences across conditions.

Three different classifiers were examined corresponding to the main stimulus effects (*SNR*: clean vs. noise classification; *ST*: 0ST vs. 4ST classification) as well as all four stimulus conditions (i.e., clean_0ST, clean_4ST, noise_0ST, noise_4ST). Chance level for classifying SNR and ST differences from the ERPs is 50% (i.e., a binary guess), whereas chance level for classifying the entire stimulus set (four options) is 25%. Since we can expect to obtain these levels of performance by chance alone, we used permutation tests to identify segments along each classifier's time course where performance was significantly above chance (one-sample *t*-test against a null of either 50% or 25%; $N = 1000$ resamples, $p < 0.05$).

2.3. *Behavioral data analysis*

2.3.1. *Identification accuracy and the "F0 benefit"*

Behavioral identification accuracy was computed for each listener as the percent of trials they labeled *both* vowel sounds 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 identification accuracy (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992)—the so-called "F0-benefit" (Arehart et al., 1997; 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 to concurrent speech segregation with and without noise interference.

2.3.2. *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. Paralleling our previous studies on the neural correlates of speech perception (e.g., Bidelman et al., 2013; Bidelman and Alain, 2015a, b; Bidelman and Walker, 2017), RTs shorter than 250 ms or exceeding 6000 ms were discarded as implausibly fast responses and lapses of attention, respectively.

2.4. *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 listeners' behavioral speech segregation, we performed weighted

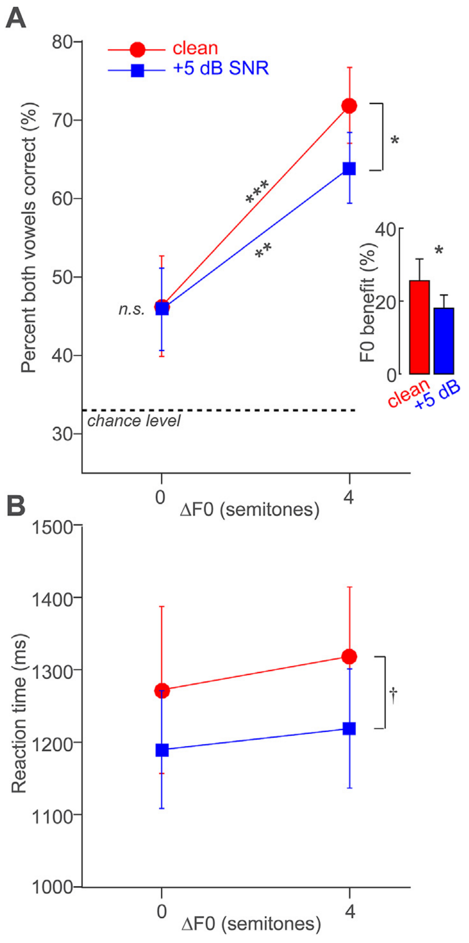

Fig. 1. Behavioral responses for segregating double-vowel stimuli. (**A**) Accuracy for identifying both tokens of double vowel mixtures. Performance is poorer when concurrent vowels contain the same F0 (0ST) and improves ~30% when they contain differing F0s (4ST). (*inset*) Behavioral F0-benefit, defined 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 speech, indicating that listeners are less successful at exploiting pitch cues when segregating acoustically-degraded signals. (**B**) Speed (i.e., RTs) of double-vowel identification. Listeners are marginally faster at identifying speech in noise. Faster RTs at the expense of poorer accuracy (panel A) suggests a time-accuracy tradeoff in double-vowel identification. Errorbars = ±1 s.e.m.; †$p \leq 0.06$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.0001$.

least square regression between their ERP amplitudes and perceptual identification accuracy and in the double-vowel task. Robust 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 (i.e., $ERP_{4ST} - ERP_{0ST}$). As in the behavioral F0 benefit, this neural analogue was computed separately for the clean and noise conditions. This neural measure was then regressed against each listener's *behavioral* F0-benefit (i.e., $PC_{4ST} - PC_{0ST}$). Paralleling our previous work on speech perception (cf. Bidelman, 2017; Bidelman and Walker, 2017), we reasoned that larger neural differentiation between 0ST and 4ST responses would correspond to larger gains in behavioral performance (i.e., larger perceptual F0-benefit). Repeating this analysis for each wave of the ERP allowed us to evaluate the earliest time at which the magnitude of neural activity mapped to behavior.

3. Results

3.1. Behavioral data

Behavioral speech identification accuracy and RTs for double-vowel segregation are shown in Fig. 1. Listeners obtained near-ceiling performance (96.9± 1.4%) when identifying single vowels in isolation. 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 × ST 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 separation (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 were 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 of both vowels corresponded with slower decision times.

3.2. ERP results

Grand averaged ERPs and scalp topographies are shown for each SNR x ST condition in Fig. 2. Scalp maps confirmed that evoked responses to speech mixtures were maximal over frontocentral electrodes, consistent with neural generators in the supratemporal plane (Alain et al., 2007; Picton et al., 1999). Visual inspection of the ERPs indicated that additive noise delayed the neural encoding of speech (i.e., $latency_{clean} < latency_{noise}$) with slightly stronger responses for double-vowels at 0STs compared to those with 4ST F0 separation.

TANOVAs (Murray et al., 2008) conducted on the ERP topographies confirmed significant modulations in evoked activity with changes in both the SNR and F0 separation of concurrent speech sounds, as well as segments sensitive to both acoustic factors (i.e., SNR × ST interaction). SNR modulated activity across nearly the entirety of the response time course (Fig. 3B). The main effect of pitch by itself was less pervasive, but significant modulations were identified in the TANOVA at a latency of ~400 ms (Fig. 3C). The

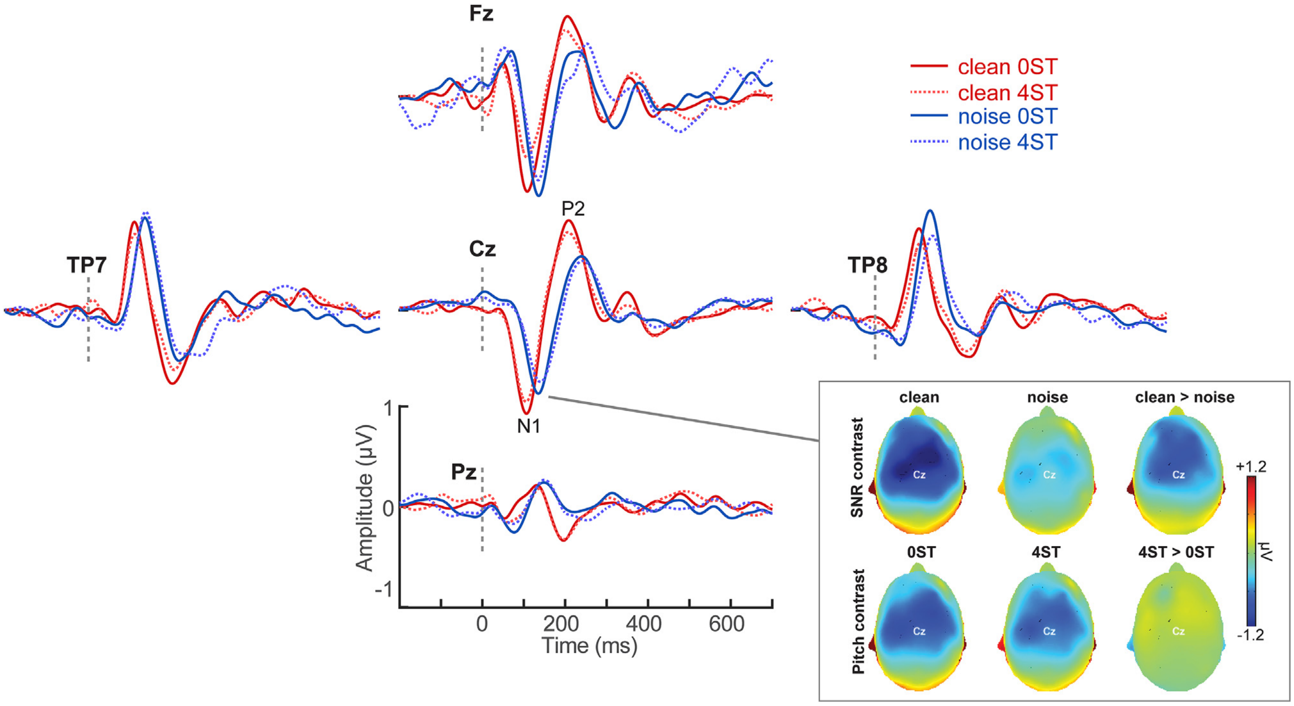

Fig. 2. Neuroelectric brain responses to concurrent vowel mixtures varying in pitch and SNR cues. (*inset*) The frontocentral distribution of scalp maps and polarity reversal near the mastoids (TP7/8) is consistent with generators in the supratemporal plane (Alain et al., 2007; Picton et al., 1999). Difference maps contrast the main stimulus effects (i.e., *SNR*: clean vs. noise; *pitch:* 4ST vs. 0 ST). Noise delays the neural encoding of concurrent speech and ERPs are slightly more robust for double-vowels containing 0ST compared to 4ST F0 separations. Positive voltage is plotted up.

significant SNR × ST interaction was circumscribed to and early (~150 ms) and late (~625 ms) time window after stimulus onset (Fig. 3D).

To further quantify these noise- and pitch-related effects, peak amplitudes and latencies were extracted for the N1, P2, and LW components identified in the initial TANOVA analysis (Fig. 4). ANOVAs conducted on N1 [$F_{1,12} = 4.83$, $p = 0.048$] and LW [$F_{1,12} = 16.90$, $p = 0.0014$] wave amplitudes showed a main effect of SNR with clean speech eliciting more robust responses than noise-degraded speech (Fig. 4A and E). P2 amplitude showed only a marginal main effect of SNR [$F_{1,12} = 4.60$, $p = 0.053$] (Fig. 4C).

Stronger effects were observed in ERP latency. N1 and LW were both earlier for clean relative to noise-degraded speech [N1: $F_{1,12} = 37.86$, $p < 0.001$; LW: $F_{1,12} = 7.30$, $p = 0.0192$]. Notably, P2 latency showed a SNR × ST interaction [$F_{1,12} = 6.57$, $p = 0.025$], paralleling the pitch × noise interaction observed behaviorally (see Fig. 1A). Post hoc contrasts revealed similar P2 latencies between clean and noisy speech in the 0ST condition (i.e., $\text{latency}_{\text{noise}} = \text{latency}_{\text{clean}}$) that diverged at 4ST (i.e., $\text{latency}_{\text{noise}} > \text{latency}_{\text{clean}}$). These results suggest that while multiple components of the ERPs are affected by SNR and pitch differences in concurrent speech mixtures, only P2 latency paralleled the SNR × ST interaction pattern observed at the behavioral level.

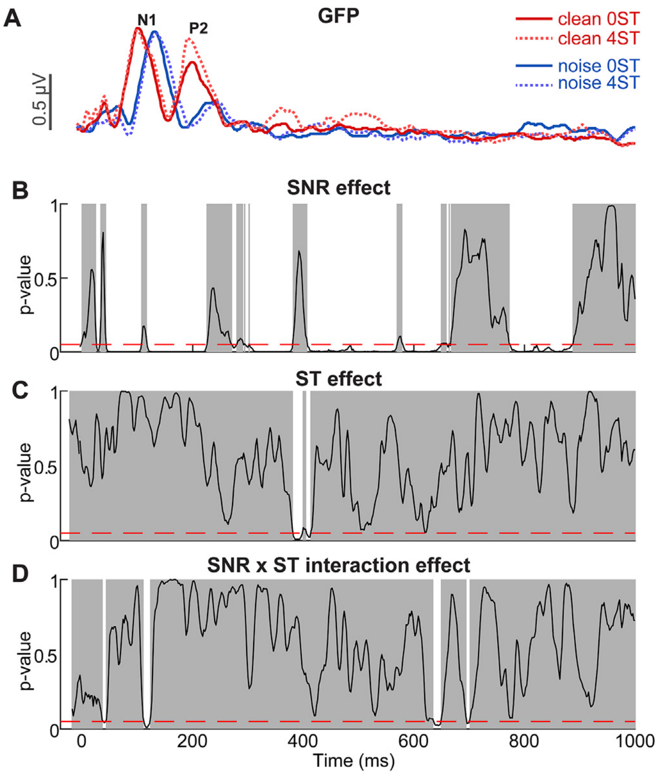

Fig. 3. Topographic ANOVA (TANOVA) revealing the time course at which the brain distinguishes concurrent speech based on SNR and pitch cues. (**A**) Mean global field power (GFP) (Lehmann and Skrandies, 1980) for each stimulus condition, quantifying the overall activation at each time sample from the aggregate multichannel evoked response. (**B**) Time course for the main effect of SNR [i.e., mean(clean_{0ST}, clean_{4ST}) − mean(noise_{0ST}, noise_{4ST})]. The trace represents the running *p*-value for the effect, computed via permutation resampling (N = 1000 shuffles; Murray et al., 2008). Dotted lines mark the $p = 0.05$ significance level (uncorrected); gray shaded areas show non-significant time segments. Clean speech is highly distinguishable from noise-degraded speech across nearly the entire epoch, but first appears in the 50–200 ms time window. (**C**) Time course for the main effect of pitch [i.e., mean($4ST_{\text{clean}}$, $4ST_{\text{noise}}$) − mean($0ST_{\text{clean}}$, $0ST_{\text{noise}}$)]. Significant modulations in the ERPs coding pitch differences are observed ~400 ms post stimulus onset. (**D**) SNR × ST interaction effect. Pitch and noise interact during the neural encoding of speech at both early (~150 ms) and late (~625 ms) time points.

3.3. Brain-behavior relationships

Associations between neural and behavioral F0-benefit are shown in Fig. 5. Of the various ERP deflections, only changes in P2 amplitude with F0-cues (4ST-0ST) were correlated with behavioral F0-benefit, particularly in noise [$r = 0.49$, $p = 0.043$]. That is, larger neural F0-benefit (reflected in P2) was associated with more accurate behavior identification performance and success in exploiting pitch cues for segregation. No other individual component showed a reliable correspondence with behavior (all *ps* > 0.05). While suggestive, some caution is still warranted when interpreting the strength of this P2 correlation because it did not survive adjustment when we corrected for multiple comparisons (i.e., $p < 0.17$; Holm, 1979). Nevertheless, ERP results indicate that pattern of responses observed in perceptual segregation (i.e., SNR × ST interaction) and behavioral F0-benefit experienced by listeners is best reflected in the latency and amplitude characteristics of the P2 wave, ~200 ms after the onset of double vowel mixtures.

We found no correlations between listeners' years of musical training and their behavioral F0- benefit nor neural F0-benefit (all *ps* > 0.53). However, this might be expected given the relatively limited musical experience in our cohort (~5 years) compared to other studies on musicianship and speech in noise listening (Bidelman and Krishnan, 2010; Parbery-Clark et al., 2009b; Zendel and Alain, 2012).

Listeners' QuickSIN scores were low (−0.73 ± 1.3 dB), consistent with the average SIN perception abilities of normal hearing listeners (i.e., 0 dB SNR loss). We found that QuickSIN scores were negatively correlated with behavioral F0 benefit in noise [$r = -0.57$, $p = 0.046$]. That is, larger gains in our double-vowel task with the introduction of pitch cues were associated with lower (i.e., better) scores on the QuickSIN. This link suggests that listeners who were better at exploiting F0-based cues for segregating isolated speech mixtures achieved better sentence-level recognition of speech in noise.

3.4. Neural classifier results

Having established that auditory neural encoding is affected by both the noise (SNR) and F0-pitch separation (ST) of concurrent vowel sounds, we next aimed to characterize how brain activity differentiated speech *spatiotemporally*. Fig. 6 shows the output of the LDA classifier, applied to discriminate the four stimulus conditions (clean/noise @ 0/4ST) at each electrode location on the scalp. For an initial analysis examining the *spatial* topography of the neural classier (temporal properties are examined below), we extracted the maximum classification accuracy achieved at each electrode location over the entire time course of the epoch. Given that we can expect to correctly classify responses 25% of the time by chance alone (i.e., guess rate for our four stimulus classes), the classifier's performance (>30–35%) is considered reliable. More importantly, the topographic distribution of the classifier corroborated our TANOVA (Figs. 2–3) and peak analyses (Fig. 4), by demonstrating that concurrent speech stimuli were optimally distinguished at frontocentral scalp sites.

To parse the *temporal* evolution of the brain's differentiation of concurrent speech based on different acoustic cues, we examined individual time courses of three neural classifiers built to discriminant SNR, ST, and the overall stimulus set (Fig. 7). Time courses were extracted in the frontocentral ROI (Fp1, Fpz, Fp2, F1, Fz, F2). This cluster was selected to parallel the ERP peak analyses and because neural classification was spatially distributed over these frontocentral regions of the scalp (i.e., Fig. 6).

Separate classifiers are shown contrasting speech-ERPs based

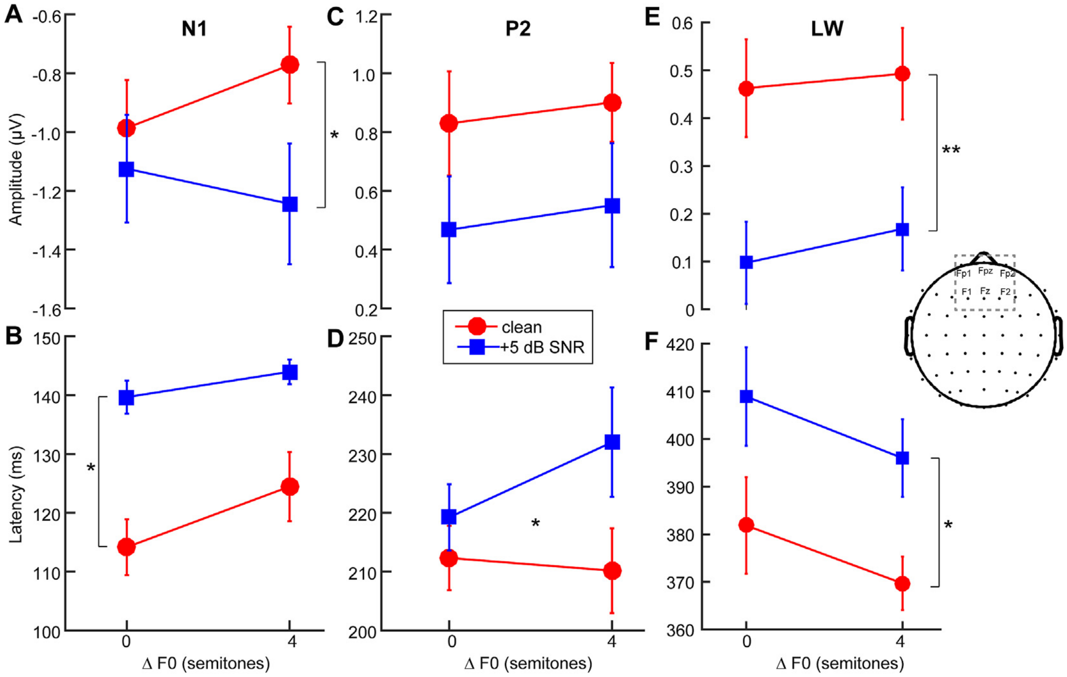

Fig. 4. ERP amplitudes and latencies are modulated by pitch and SNR differences in concurrent speech mixtures. Amplitude and latency data for the N1 (**A-B**), P2 (**C-D**), and LW (**E-F**) waves extracted from a ROI cluster of six frontocentral electrodes (Fp1, Fpz, Fp2, F1, Fz, F2) (*inset*). While most responses were stronger and earlier for clean relative to noise-degraded speech, only P2 latency showed a SNR × ST interaction that paralleled the interaction pattern observed behaviorally (cf. Fig. 1A). $^*p < 0.05$, $^{**}p < 0.01$.

on SNR (blue; clean vs. noise), number of STs between vowel F0s (red; 0ST vs. 4ST), and all four stimulus conditions (black; clean/noise @ 0/4ST). Chance levels for the binary contrasts of SNR and ST are 50%, whereas chance for classifying the entire stimulus set is 25% (dotted lines). We found that neural responses segregated clean from noise-degraded speech as early as ~100 and 200 ms, aligning with the N1 and P2 waves of the ERP. Classification of speech based solely on F0 cues was weaker but occurred briefly at ~700 ms (see also Fig. 3D). Overall classification performance discriminating the four stimulus classes (2 SNR x 2 ST) was also well

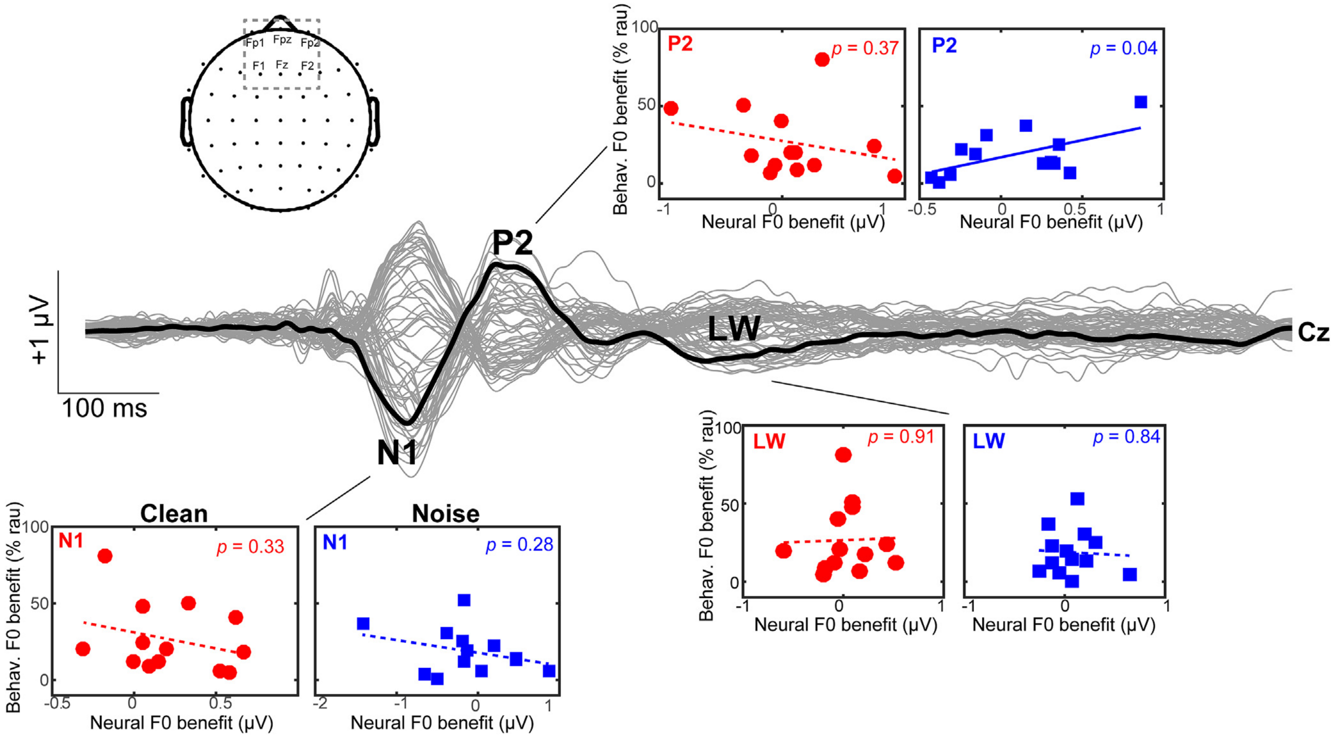

Fig. 5. Brain-behavior relations underlying double-vowel segregation. Individual panels show correlations between listeners' *behavioral* (see Fig. 1A) and *neural* F0-benefit. Larger F0-benefit reflects more successful neural/behavioral speech segregation with the addition of pitch cues. Although neural encoding of concurrent speech is heavily modulated by SNR and ST individually (see Fig. 4), only changes in P2 amplitude (across subjects) are correlated with behavior. Larger neural differentiation of speech (in noise) is associated with more accurate behavioral identification. Bold trace, potential at Cz. Sold lines, significant relations; dotted lines, *n.s.* relations.

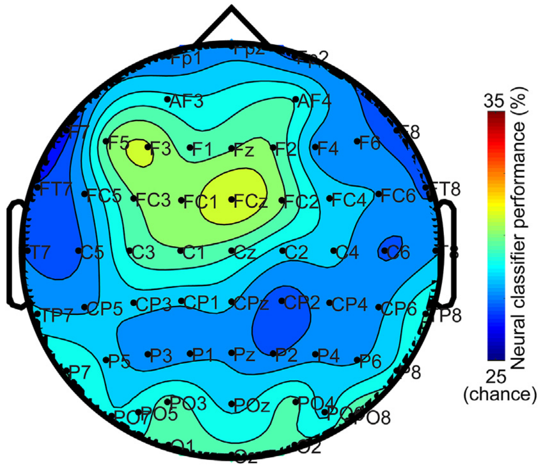

Fig. 6. Grand average topography of neural classification. For each electrode, we plot the maximum neural classification accuracy (% correct) across the time course of the ERPs. Chance level is 25% for the four option stimulus set (i.e., clean_0ST, clean_4ST, noise_0ST, noise_4ST). Cool colors denote scalp locations where evoked activity poorly classifies different speech conditions; hot colors, locations showing above chance discriminability. Concurrent speech stimuli are optimally distinguished at frontocentral scalp sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

above 25% chance and occurred within 250 ms of stimulus onset, during the N1-P2 complex of evoked responses.

To evaluate the behavioral relevance of our neural classification results, we examined correlations between each classifier's performance metrics and listeners' corresponding perceptual segregation. For the SNR and ST time courses (i.e., Fig. 7), we measured peak accuracy and latency at which the classifier achieved maximal segregation of the stimulus conditions. Search windows for peak quantification were guided by the permutation analysis, which identified temporal segments that showed reliable (above chance) discrimination of speech conditions (i.e., 0–200 ms for the SNR classifier; 600–800 ms for the ST classifier; see Fig. 7). We then regressed the classifiers' max accuracy and latency against listeners' corresponding change in behavioral accuracy for each of the SNR and ST speech contrasts. For instance, for the effect of SNR, we ask if the latency/max accuracy of the brain's differentiation of clean vs. noise-degraded speech corresponds with a perceptual improvement in identifying clean vs. noisy speech. For ST-based segregation, we ask if the latency/max accuracy of neural differentiation of 4 ST vs. 0 ST speech corresponds to the perceptual improvement in identifying speech mixtures with and without pitch cues.

Correlations between classifier performance and behavior are shown in Fig. 8. For the SNR classifier, we found that the latency of peak neural segregation of clean vs. noisy speech was negatively correlated with behavioral segregation accuracy [$r = -0.69$, $p = 0.009$]. That is, earlier (more efficient) neural speech differentiation predicted a larger improvement in perceptual identification for clean relative to noise-degraded vowels. Similarly, peak neural classification accuracy was positively correlated with listeners' ability to exploit pitch cues for segregation [$r = 0.57$, $p = 0.044$]. That is, larger neural differentiation between ST conditions was associated with better behavioral speech segregation using pitch cues. These results help clarify the behavioral relevance of our neural classifier analysis by demonstrating that the perceptual segregation of speech mixtures is determined by how accurately and how early brain activity distinguishes concurrent speech tokens.

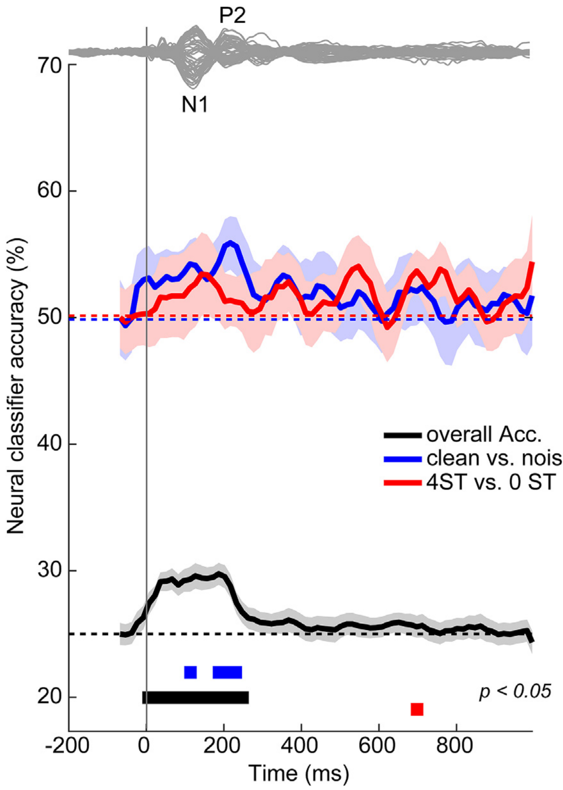

Fig. 7. Time course for distinguishing concurrent speech from neural responses. Results reflect physiological speech differentiation at the frontocentral ROI (Fp1, Fpz, Fp2, F1, Fz, F2). Separate classifiers are shown contrasting ERPs based on SNR (blue; clean vs. noise), number of STs between vowel F0s (red; 0ST vs. 4ST), and all four stimulus conditions (black; clean/noise @ 0/4ST). Chance level for the binary contrasts of SNR and ST are 50%; chance level for overall classification of the four stimulus classes is 25% (dotted lines). Bars (■) below traces show time segments where each classifier performs significantly above chance (permutation test; $p < 0.05$, $N = 1000$ resamples). Neural responses segregate clean from noise-degraded speech at ~100 and 200 ms (i.e., N1 and P2 waves). Classification of speech with and without F0 cues is weaker but occurs briefly at ~700 ms (see also Fig. 3D). Overall classification of the entire stimulus set (2 SNR x 2 ST) is also well above chance and occurs within 250 ms of stimulus onset (i.e., during the N1-P2 complex). Shading = ±1 s.e.m. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

In the present study, we examined neuroelectric brain activity as listeners rapidly identified double vowel stimuli varying in their pitch (F0) and noise level (SNR) to better delineate the time course of speech segregation based on intrinsic and extrinsic acoustic factors. Consistent with the large body of literature on double vowel identification (e.g., Alain et al., 2007, 2017; Assmann and Summerfield, 1989; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016), results confirm that listeners can exploit F0 differences between vowels to segregate speech. We extend these previous studies but demonstrating that “F0-benefit” (i.e., improvement in accuracy with pitch cues) interacts with the degree of noise interference in the acoustic environment. Perceptual F0-benefit was larger for clean compared to noise-degraded (+5 dB SNR) speech. These behavioral data were paralleled in several early and late modulations in the ERPs which reflected different properties of parsing concurrent speech mixtures.

Our electrophysiological data corroborate recent studies on the neural correlates of concurrent speech segregation by demonstrating early modulations in the ERPs within ~250 ms of sound onset that carry information on multiple auditory objects (Alain

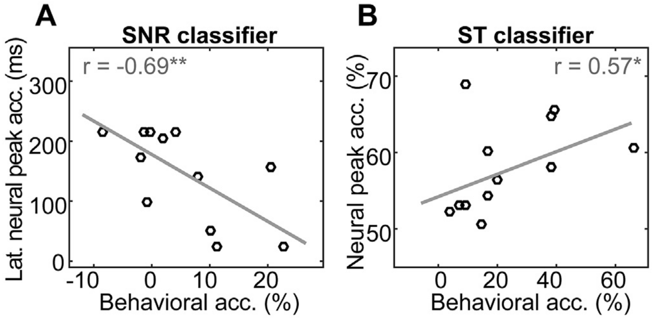

Fig. 8. Relations between neural classifier metrics and behavioral speech segregation performance. (**A**) Latency of peak neural classification (0–200 ms time window, see blue trace, Fig. 7) in distinguishing clean and noise speech ERPs is negatively correlated with behavioral segregation accuracy. More efficient neural differentiation of speech is associated with a larger improvement in perceptual identification for clean relative to noise-degraded vowel mixtures. (**B**) Peak neural classification accuracy is positively correlated with listeners' ability to exploit pitch cues for segregation. Larger neural differentiation of speech between ST conditions is associated with better behavioral segregation. $^{*}p < 0.05$, $^{**}p < 0.01$.

et al., 2007, 2017). We observed rapid transient changes in the N1 and P2 dependent on the F0 separation of simultaneous vowel sounds and the SNR of a concurrent noise masker. Relative to clean speech, evoked responses in noise were generally delayed and weaker, consistent with noise-related changes observed in the encoding of isolated speech sounds (Bidelman and Howell, 2016; Billings et al., 2010, 2013). Some studies have suggested that low-level intensity background noise can actually enhance cortical responses to auditory stimuli (Alain et al., 2012, 2014, 2009; Bidelman and Howell, 2016; Parbery-Clark et al., 2011), a facilitation that may reflect engagement of the "antimasking" function of the peripheral efferent system (Alain et al., 2009; Bidelman and Bhagat, 2015; Winslow and Sachs, 1987). However, the lack of noise-related facilitation in the current study compared to others is likely due to our weaker noise level and the use of double vowel mixtures compared to isolated speech. Noise-related changes in the speech ERPs are thought to reflect a reduction in neural synchrony with increasing levels of noise (Ponjavic-Conte et al., 2013). This explanation could account for the attenuated responses we find for noisier speech mixtures compared to their clean counterparts.

In contrast to the more pervasive SNR-related ERP modulations which lasted over the entirety of the response (e.g., Fig. 3B), F0-related changes in double-vowel encoding were more muted and circumscribed to the ~400 ms time window (e.g., Fig. 3C). This later modulation is consistent with previous neuroimaging studies which have observed a similar deflection 350–400 ms after sound onset that covaries with the F0-difference between vowels (Alain et al., 2005a, 2007; Reinke et al., 2003). That this late wave is linked to segregation processing is suggested by the fact that more successful learners in double-vowel tasks show enhancements in later sustained activity 300–400 ms after speech onset (Alain et al., 2007, 2015).

Our neural classifier analysis helps elucidate the time course of these extrinsic (noise) and intrinsic (pitch) acoustic stressors on concurrent speech segregation. We found that brain activity reliably distinguished speech with and without external noise within 150–250 ms after stimulus onset. In contrast to SNR-based segregation, the neural differentiation of speech mixtures based on pitch cues alone took considerably longer (~700 ms) (Fig. 7). Thus, corroborating results from both our TANOVA and time-varying neural classification analyses imply early and late processes that unfold during the parsing of simultaneous speech. Consistent with multistage models of concurrent speech segregation (Alain et al., 2005a), we argue that the early process reflects pre-perceptual, sensory-based segregation occurring in our near auditory cortex that tags the acoustic clarity of speech (i.e., clean vs. noise) within 200 ms (Bidelman and Howell, 2016). The frontocentral topography of both the raw ERPs as well as neural classifier accuracy (Fig. 7) is consistent with neural generators in the supratemporal plane (Alain et al., 2007; Picton et al., 1999). In contrast, the later activation (~400 ms and beyond) could index post-perceptual processes, reflecting the match of each vowel constituent to their respective memory templates (Alain et al., 2007). Evidence that this later component reflects cognitive processing is supported by studies showing similar late (~400 ms) activation that occurs only during active (but not passive) concurrent sound segregation (Alain et al., 2001; Bidelman and Alain, 2015a). Compared to SNR-based segregation, F0-based segregation is arguably more difficult and cognitively demanding than parsing clean vs. noise-degraded speech. Thus, the later time course of the ST compared to SNR effect in our data (e.g., Figs. 3C and 7) could reflect the higher cognitive demand and/or perceptual confusion when segregating speech mixtures using pitch cues alone.

While our neural classifier and TANOVA results suggest that neural activity can adequately differentiate speech based on either SNR or pitch cues, these analyses do not address the *behavioral relevance* of neural activity to perceptual sound segregation. In this regard, brain-behavioral correlations help reveal the earliest time at which neural activity maps to behavior. At the behavioral level, we found that noise and pitch interacted during listeners' double-vowel segregation (Fig. 1A). That is, while all participants experienced the classic F0-benefit (i.e., improvement in identification accuracy with pitch cues), the advantage was dependent on the amount of background noise that overlapped target speech mixtures; listeners experienced stronger F0-benefit for clean relative to noise-degraded speech. Paralleling the behavioral data, ERPs showed significant SNR x F0 modulations within the timeframe of the N1-P2 complex (Fig. 3D). However, only the P2 wave showed a similar SNR x F0 pattern as behavior (Fig. 4D) and actually correlated with listeners' perception (Fig. 5). Thus, while we observe multiple time courses to concurrent speech segregation mechanisms, perceptual success in parsing multiple streams seems driven by coding in the timeframe of P2 (150–200 ms). These results are consistent with the notion that early components like the N1 reflect exogenous properties of the acoustic signal whereas P2 further reflects the endogenous properties of the signal's identity and the recognition of perceptual objects (e.g., Alain et al., 2007; Bidelman and Lee, 2015; Bidelman and Alain, 2015b; Bidelman et al., 2013; Wood et al., 1971).

In sum, we find a dynamic time course for concurrent speech sound processing that depends on both extrinsic and intrinsic acoustic factors. The earlier timing of neural speech differentiation based on noise (SNR) compared to pitch (F0) implies that the cortical extraction of speech from extrinsic noise is more efficient than distinguishing speech using intrinsic voice pitch cues alone. Nevertheless, our findings demonstrate that noise and pitch information interact relatively early (few hundred milliseconds) in the neural hierarchy as cerebral cortex arrives at the identity of concurrent speech signals.

Acknowledgments

The authors thank Amy Fehrenbach for assistance in data collection.

References

Alain, C., Arnott, S.R., Picton, T.W., 2001. Bottom-up and top-down influences on auditory scene analysis: evidence from event-related brain potentials. J. Exp. Psychol. Hum. Percept. Perform. 27, 1072–1089.

Alain, C., McDonald, K., Van Roon, P., 2012. Effects of age and background noise on processing a mistuned harmonic in an otherwise periodic complex sound. Hear. Res. 283, 126–135.

Alain, C., Roye, A., Salloum, C., 2014. Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex. Front. Sys. Neurosci. 8, 8.

Alain, C., Snyder, J.S., He, Y., Reinke, K.S., 2007. Changes in auditory cortex parallel rapid perceptual learning. Cereb. Cortex 17, 1074–1084.

Alain, C., Quan, J., McDonald, K., Van Roon, P., 2009. Noise-induced increase in human auditory evoked neuromagnetic fields. Eur. J. Neurosci. 30, 132–142.

Alain, C., Zhu, K.D., He, Y., Ross, B., 2015. Sleep-dependent neuroplastic changes during auditory perceptual learning. Neurobiol. Learn. Mem. 118, 133–142.

Alain, C., Reinke, K., He, Y., Wang, C., Lobaugh, N., 2005a. Hearing two things at once: neurophysiological indices of speech segregation and identification. J. Cogn. 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.

Alain, C., Reinke, K., McDonald, K.L., Chau, W., Tam, F., Pacurar, A., Graham, S., 2005b. Left thalamo-cortical network implicated in successful speech separation and identification. Neuroimage 26, 592–599.

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.

Assmann, P.F., 1996. Tracking and glimpsing speech in noise: role of fundamental frequency. J. Acoust. Soc. Am. 100, 2680.

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.

Bidelman, G.M., 2015. 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., Krishnan, A., 2010. Effects of reverberation on brainstem representation of speech in musicians and non-musicians. Brain Res. 1355, 112–125.

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

Bidelman, G.M., Lee, C.-C., 2015. Effects of language experience and stimulus context on the neural organization and categorical perception of speech. Neuroimage 120, 191–200.

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., Bhagat, S.P., 2015. Right ear advantage drives the link between olivocochlear efferent “antimasking” and speech-in-noise listening benefits. NeuroReport 26, 483–487.

Bidelman, G.M., Chung, W.-L., 2015. Tone-language speakers show hemispheric specialization and differential cortical processing of contour and interval cues for pitch. Neuroscience 305, 384–392.

Bidelman, G.M., Dexter, L., 2015. Bilinguals at the "cocktail party": dissociable neural activity in auditory-linguistic brain regions reveals neurobiological basis for nonnative listeners' speech-in-noise recognition deficits. Brain Lang. 143, 32–41.

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., 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.

Bidelman, G.M., Weiss, M.W., Moreno, S., Alain, C., 2014. Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. Eur. J. Neurosci. 40, 2662–2673.

Bidelman, G.M., Lowther, J.E., Tak, S.H., Alain, C., 2017. Mild cognitive impairment is characterized by deficient hierarchical speech coding between auditory brainstem and cortex. J. Neurosci. 37, 3610–3620.

Bidet-Caulet, A., Bertrand, O., 2009. Neurophysiological mechanisms involved in auditory perceptual organization. Front. Neurosci. 3, 182–191.

Billings, C.J., Tremblay, K.L., Stecker, G.C., Tolin, W.M., 2009. Human evoked cortical activity to signal-to-noise ratio and absolute signal level. Hear. Res. 254, 15–24.

Billings, C.J., Bennett, K.O., Molis, M.R., Leek, M.R., 2010. Cortical encoding of signals in noise: effects of stimulus type and recording paradigm. Ear Hear 32, 53–60.

Billings, C.J., McMillan, G.P., Penman, T.M., Gille, S.M., 2013. Predicting perception in noise using cortical auditory evoked potentials. J. Assoc. Res. Oto 14, 891–903.

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.S., 1990. Auditory Scene Analysis MIT, Cambridge, MA.

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.

de Cheveigné, A., McAdams, S., Marin, C.M.H., 1997a. Concurrent vowel identification. II. Effects of phase, harmonicity, and task. J. Acoust. Soc. Am. 101, 2848–2856.

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

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

Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. U. S. A 111, 1–6.

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.

Helfer, K., Wilber, L., 1990. Hearing loss, aging, and speech perception in reverberation and in noise. J. Speech Hear. Res. 33, 149–155.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Irimajiri, R., Golob, E.J., Starr, A., 2005. Auditory brain-stem, middle- and long-latency evoked potentials in mild cognitive impairment. Clin. Neurophysiol. 116, 1918–1929.

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.

Koenig, T., Melie-Garcia, L., 2010. A method to determine the presence of averaged event-related fields using randomization tests. Brain Topogr. 23, 233–242.

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.

Lee, S., Bidelman, G.M., 2017. Objective identification of simulated cochlear implant settings in normal-hearing listeners via auditory cortical evoked potentials. Ear Hear. http://dx.doi.org/10.1097/AUD.0000000000000403 (in press).

Lehmann, D., Skrandies, W., 1980. Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalogr. Clin. Neurophysiol. 48, 609–621.

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

Macdonald, E.N., Pichora-Fuller, M.K., Schneider, B.A., 2010. Effects on speech intelligibility of temporal jittering and spectral smearing of the high-frequency components of speech. Hear. Res. 261, 63–66.

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

Murray, M.M., Brunet, D., Michel, C.M., 2008. Topographic ERP analyses: a step-by-step tutorial review. Brain Topogr. 20, 249–264.

Nabelek, A.K., Letowski, T.R., Tucker, F.M., 1989. Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Am. 86, 1259–1265.

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.

Oxenham, A.J., 2008. Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. Trends Amplif. 12, 316–331.

Palmer, A.R., 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.

Parbery-Clark, A., Skoe, E., Kraus, N., 2009a. Musical experience limits the degradative effects of background noise on the neural processing of sound. J. Neurosci. 29, 14100–14107.

Parbery-Clark, A., Skoe, E., Lam, C., Kraus, N., 2009b. Musician enhancement for speech-in-noise. Ear Hear 30, 653–661.

Parbery-Clark, A., Marmel, F., Bair, J., Kraus, N., 2011. What subcortical-cortical relationships tell us about processing speech in noise. Eur. J. Neurosci. 33, 549–557.

Parsons, T.W., 1976. Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Am. 60, 911–918.

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

Ponjavic-Conte, K.D., Hambrook, D.A., Pavlovic, S., Tata, M.S., 2013. Dynamics of distraction: competition among auditory streams modulates gain and disrupts inter-trial phase coherence in the human electroencephalogram. PLoS ONE 8, e53953.

Qin, M.K., Oxenham, A.J., 2005. Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. Ear Hear 26, 451–460.

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

Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34, 114–123.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Sinex, D.G., Sabes, J.H., Li, H., 2002. Responses of inferior colliculus neurons to harmonic and mistuned complex tones. Hear. Res. 168, 150–162.

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

Swaminathan, J., Heinz, M.G., 2012. Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. J. Neurosci. 32, 1747–1756.

Van Noorden, L.P.A.S., 1975. Temporal Coherence in the Perception of Tone Sequences. Eindhoven University of Technology, Eindhoven, The Netherlands. Doctoral Dissertation.

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.

Winslow, R.L., Sachs, M.B., 1987. Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. J. Neurophysiol. 57, 1002–1021.

Wood, C.C., Goff, W.R., Day, R.S., 1971. Auditory evoked potentials during speech perception. Science 173, 1248–1251.

Zendel, B.R., Alain, C., 2012. Musicians experience less age-related decline in central auditory processing. Psychol. Aging 27, 410–417.