Behavioral/Systems/Cognitive

Dynamic and Task-Dependent Encoding of Speech and Voice by Phase Reorganization of Cortical Oscillations

Milene Bonte, Giancarlo Valente, and Elia Formisano
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands

Speech and vocal sounds are at the core of human communication. Cortical processing of these sounds critically depends on behavioral demands. However, the neurocomputational mechanisms enabling this adaptive processing remain elusive. Here we examine the task-dependent reorganization of electroencephalographic responses to natural speech sounds (vowels /a/, /i/, /u/) spoken by three speakers (two female, one male) while listeners perform a one-back task on either vowel or speaker identity. We show that dynamic changes of sound-evoked responses and phase patterns of cortical oscillations in the alpha band (8–12 Hz) closely reflect the abstraction and analysis of the sounds along the task-relevant dimension. Vowel categorization leads to a significant temporal realignment of responses to the same vowel, e.g., /a/, independent of who pronounced this vowel, whereas speaker categorization leads to a significant temporal realignment of responses to the same speaker, e.g., speaker 1, independent of which vowel she/he pronounced. This transient and goal-dependent realignment of neuronal responses to physically different external events provides a robust cortical coding mechanism for forming and processing abstract representations of auditory (speech) input.

Key words: **auditory cortex; EEG; alpha; synchrony; language; speech**

Introduction

Human speech conveys linguistic content and speaker-specific voice information with crucial relevance in our daily life. Dependent on the current behavioral goal, we may choose to focus our attention on either of these types of information. Such adaptive behavior requires a computational mechanism that enables different (abstract) representations of the same acoustic input. The neural implementation of this mechanism remains unknown and is examined in the present study by analyzing electroencephalographic (EEG) responses during task-dependent categorization of the same speech stimuli into phonemes or speakers.

Cortical processing of voice (Belin et al., 2000) and speech (Binder et al., 2000; Scott et al., 2000) may rely on distinct systems in the superior temporal cortex. The temporal dynamics of these systems has been studied with EEG and magnetoencephalography (MEG), mostly by separately addressing the two dimensions. Preattentive discrimination of voices (Titova and Näätänen, 2001; Beauchemin et al., 2006) and the extraction of acoustic–phonetic speech features (Näätänen et al., 1997; Poeppel et al., 1997; Obleser et al., 2004b) unfold within 100–200 ms. The product of these computations may lead to the formation of intermediate representations, such as phonemes, that are invariant to changes in the acoustic input and that can be used for further linguistic processing (McClelland and Elman, 1986; Norris and McQueen, 2008). A plausible neural mechanism that enables this abstraction, however, has not been demonstrated.

Here we examine EEG responses elicited by three phonemes (vowels) spoken by three speakers (Fig. 1), while listeners performed one-back tasks on either vowel or speaker identity (vowel and speaker task). Because they may reflect complementary aspects of neural processing (Makeig et al., 2002; Klimesch et al., 2007b), we report both event-related potential (ERP) and oscillatory responses. We hypothesize that the task-dependent categorization of speech sounds into vowels or speakers relies on distinctive phase patterns of cortical oscillations. This hypothesis is based on findings in several areas of research. Animal research indicates that cortical oscillations are at the basis of object perception and mediate selective enhancement of relevant stimulus information (Engel et al., 2001; Salinas and Sejnowski, 2001). Theoretical work suggests that transient stimulus evoked oscillations are robust to irrelevant acoustic variations in the speech signal and may provide a neural mechanism for speech perception (Hopfield and Brody, 2001). Furthermore, in humans, phase patterns of cortical oscillations have been reported to discriminate the acoustic structure of spoken sentences (Ahissar et al., 2001; Luo and Poeppel, 2007). According to our hypothesized neural coding scheme (Fig. 2), task-dependent top-down processes trigger a transient phase reorganization. We test this hypothesis by computing two intertrial-phase-coherence (ITC) measures from the EEG responses to our nine speech stimuli: (1) the average ITC for each of the vowels, regardless of speakers (ITC-vowel), and (2) the average ITC for each of the speakers, regardless of vowels (ITC-speaker). We predict that the vowel task aligns phases of neural responses to vowels regardless of speakers (ITC-vowel > ITC-speaker) and the speaker task aligns

Received Aug. 5, 2008; revised Jan. 5, 2009; accepted Jan. 7, 2009.

This work was supported by the Netherlands Organization for Scientific Research Innovative Research Incentives Scheme VENI Grant 451-07-002 (M.B.) and VIDI Grant 452-04-330 (E.F.). We thank Ruben Janssen for assistance in EEG data acquisition.

Correspondence should be addressed to Milene Bonte, Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: m.bonte@psychology.unimaas.nl.

DOI:10.1523/JNEUROSCI.3694-08.2009

 0270-6474/09/291699-08$15.00/0

phases of neural responses to speakers regardless of vowels (ITC-speaker > ITC-vowel).

Materials and Methods

Participants. Fourteen healthy Dutch-speaking undergraduate students (8 females; 1 left handed) participated in the study. None of the participants had a history of hearing loss or neurological abnormalities. Participants gave their informed consent and received course credits or payment for participation. Approval for the study was granted by the Ethical Committee of the Faculty of Psychology at the University of Maastricht.

Stimuli. Stimuli were speech sounds consisting of three natural Dutch vowels (/a/, /i/, and /u/) spoken by three native Dutch speakers (sp1: female, sp2: male, and sp3: female). To introduce acoustic variability typical of natural speech perception, for each vowel and for each speaker we included three different tokens. For instance, condition "a-sp1" included three different utterances of the vowel /a/ spoken by speaker 1 (Fig. 1). Stimuli were digitized at a sampling rate of 44.1 kHz, D/A converted with 16 bit resolution, bandpass filtered (80 Hz to 10.5 kHz), downsampled to 22.05 kHz, and edited with PRAAT software (Boersma and Weenink, 2002). Stimulus length was equalized to 230 ms (original range 172–338 ms), by using PSOLA (100–300 Hz as extrema for the F0 contour). We carefully checked our stimuli for possible alterations in F0 after length equalization and did not find any detectable changes. Sound intensity level was numerically equalized across stimuli by matching RMS values. To avoid acoustic transients (clicks) that would be created by a sharp cutoff, stimuli were faded with 50 ms linear onset and offset ramps.

Experimental design and procedure. We investigated task dependent processing of our stimuli by comparing the processing of the nine speech conditions (a-sp1, a-sp2, a-sp3, i-sp1, i-sp2, i-sp3, u-sp1, u-sp2, u-sp3) during the performance of two different tasks: (1) speaker task and (2) vowel task. We also included a passive control condition (passive task), in which subjects were instructed to listen to a random sequence of speech stimuli while fixating a fixation cross in the center of a computer monitor. The speaker task involved a one-back task on speaker identity, i.e., subjects were instructed to press the space bar on the computer keyboard whenever the same speaker was repeated. During the vowel task, subjects performed a one-back task on vowel identity. Because our task manipulation served to trigger the processing of the same stimuli in a speaker versus vowel modus and we did not aim to investigate response-related activity, targets only occurred in 6.5% of the trials. Trials including targets and/or button presses (correct responses, omissions, and false positives) were not included in the EEG analysis. During the speaker and vowel task the words "Person" and "Vowel" were presented in the center of the computer monitor respectively.

Each of the three tasks involved two blocks and a total of 450 nontarget trials (50 trials for each of the 9 speech conditions). Speaker and vowel task blocks additionally included 16 target trials (total 32 targets per task). Stimuli were presented binaurally through loudspeakers at a comfortable listening level. Stimulus onset asynchrony (SOA) for two consecutive stimuli randomly varied between 3.0 and 3.5 s.

All subjects participated in two EEG sessions during which they performed (1) two passive task blocks followed by two speaker task blocks and (2) two passive task blocks followed by two vowel task blocks. The order of the two sessions was counterbalanced between subjects with a between session break of minimally 1 week (maximally 4 weeks). Only the passive blocks of the first session were included in the analysis. Before the speaker and vowel tasks, subjects performed a practice session to make sure the tasks were understood and performed accurately.

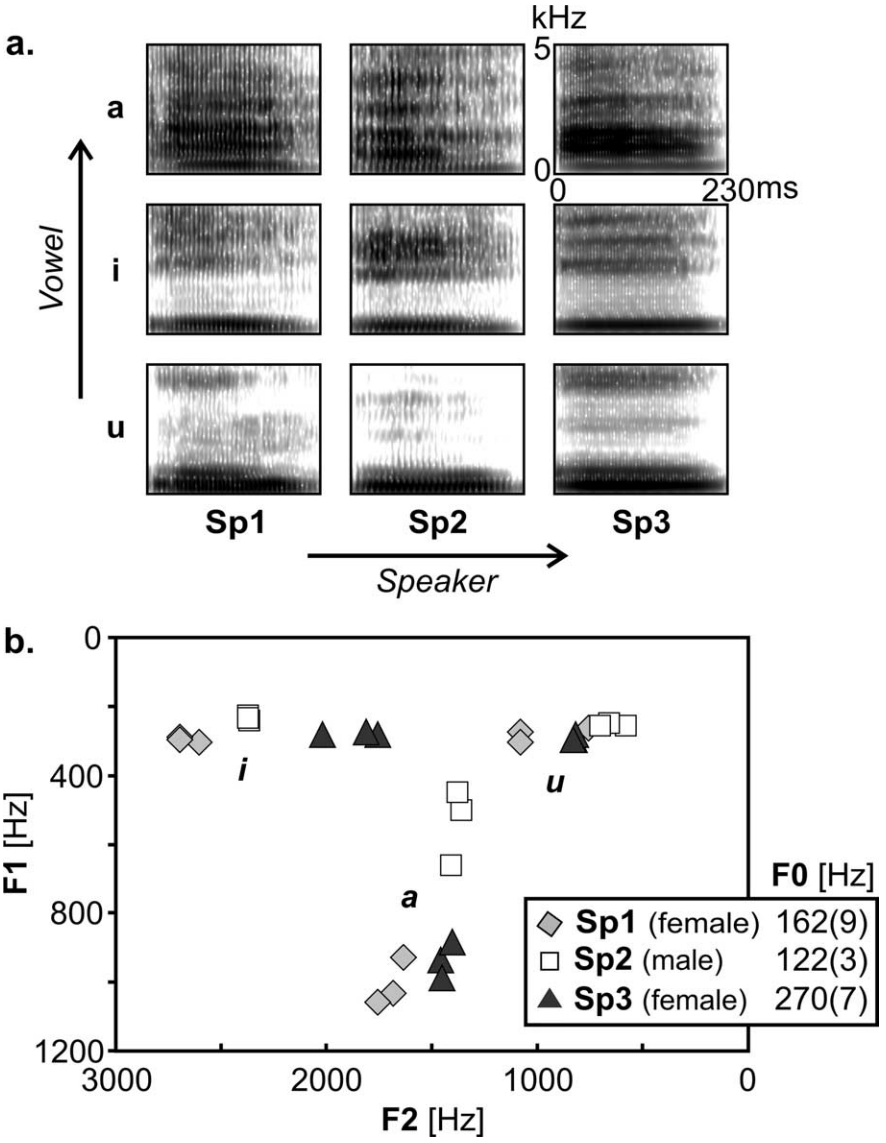

Figure 1. Stimuli and design. ***a***, Spectrograms of one exemplar of each of the nine speech conditions. Stimuli consisted of three vowels (/a/, /i/, /u/) pronounced by three speakers (speaker 1, 2, 3). ***b***, F1/F2 formant values for all stimuli (3 utterances per vowel for each speaker) and mean ± SD pitch (F0) values for each of the three speakers.

Data recording and analysis. EEG data were recorded (0.01–100 Hz, sampling rate 250 Hz) in a sound-attenuating and electrically shielded room from 61 equidistant electrode positions (Easycap, montage no. 10) relative to a left mastoid reference. An additional EOG electrode was placed below the left eye. All electrode impedance levels (EEG and EOG) were kept at <5 kΩ. Data were analyzed using the EEGLAB toolbox (Delorme and Makeig, 2004) and custom Matlab scripts. Raw EEG data were bandpass filtered (0.5–70 Hz), epoched from -1.0 to 1.5 s relative to stimulus onset, and baseline corrected (1 s prestimulus interval). Removal of artifacts was performed in two steps. First, the data were visually inspected and epochs containing nonstereotypical artifacts including high-amplitude, high-frequency muscle noise, swallowing, and electrode cable movements, were rejected. Second, stereotypical artifacts, including eye movements, eye blinks, and heart beat artifacts, were corrected

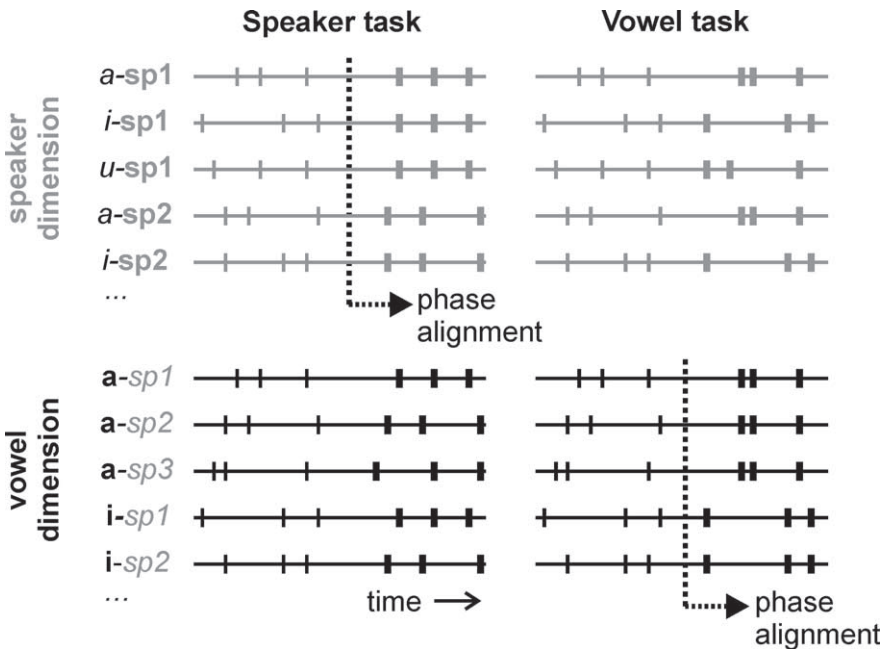

Figure 2. Hypothesized neural coding scheme. The hypothesized mechanism relies on transient phase reorganizations of stimulus-locked oscillatory responses. During initial stimulus-driven analysis each speech stimulus (vowels /a/, /i/, /u/; spoken by speakers sp1, sp2, sp3) is assumed to be characterized by a unique pattern of neural responses (as indicated by tick marks). In a subsequent time window, top-down task demands lead to a phase alignment of cortical oscillations such that neural responses become more similar along the task-relevant stimulus dimension. Thus, the speaker task aligns phases for each of the speakers, independent of the vowel that was pronounced (left part, top row), whereas the vowel task aligns phases for each of the vowels, independent of who was speaking (right part, bottom row). Note that the top and bottom rows are schematic representations of the same stimulus-evoked neural responses, only their arrangement changes (along speaker dimension in the top row; along vowel dimension in the bottom row).

with extended INFOMAX ICA (Lee et al., 1999) as implemented in EEGLAB. ICA was performed separately for each task (speaker, vowel, passive), and each subject using function runica with default extended-mode training parameters and a stopping weight change of 1×10^{-7} (Delorme and Makeig, 2004). Because data were recorded at 62 channels (61 EEG electrodes, 1 EOG electrode), runica decomposed the data in 62 component activations per task per subject. These component activations were categorized as EEG activity or nonbrain artifacts by visual inspection of their scalp topographies, time courses, and frequency spectra. Criteria for categorizing component activations as EEG activity included (1) a scalp topography consistent with an underlying dipolar source, (2) spectral peak(s) at typical EEG frequencies, and (3) regular responses across single trials, i.e., an EEG response should not occur in only a few trials (Delorme et al., 2004). Based on these criteria, component activations representing nonbrain artifacts were removed, and EEG data were reconstructed from the remaining component activations representing brain activity. The reconstructed data were based on a mean (SD) number of 23 (±4) EEG components per subject for the vowel task, 24 (±4) components for the speaker task, and 20 (±4) components for the passive task. These data were baseline corrected (1 s prestimulus interval) and used for further ERP and time–frequency analyses.

ERP measures included mean amplitudes in the N1 (90–130 ms), P2 (170–230 ms), and P340 (310–370 ms) windows. Because an initial analysis of ERP data did not show effects of ERP latency, latency measures were not included in the analysis. Amplitude measures of the N1, P2, and P340 responses to each of our nine speech conditions were analyzed with a repeated-measures ANOVA with task (speaker vs vowel task), vowel (/a/, /i/, /u/), and speaker (sp1, sp2, sp3) as within-subjects factors. ERP data of the passive task were not included in our ANOVA because the present study was designed to compare the active modulation of speech processing during the vowel versus speaker tasks and the passive task merely served as a control condition. Furthermore, EEG responses elicited during the passive task turned out to resemble those elicited during the vowel task, and its inclusion would thus reduce the sensitivity of our analysis with respect to our task modulation of interest. ANOVA results are reported using thresholded ($p < 0.05$) *F*-maps in which all electrodes were included (see Fig. 3; supplemental Fig. 2, available at www.jneurosci.org as supplemental material).

The ANOVA did not show any significant task-by-vowel-by-speaker interaction. We thus considered the interaction between responses elicited by our nine speech conditions in the two active tasks. Focusing on this interaction allowed us eliminating ERP differences due to the physical differences between the sounds as well as overall task effects, and thereby targeting a more abstract level of stimulus representation. To illustrate the ERP effects (see Fig. 3), we constructed two types of event-related averages reflecting the two possible stimulus dimensions: (1) "vowel " averages, i.e., averages of the ERP responses to the three vowels regardless of speakers [e.g., ERP /a/ = (ERP a-sp1 + ERP a-sp2 + ERP a-sp3)/3], and (2) "speaker" averages, i.e., averages of the ERP responses to the three speakers regardless of vowels [e.g., ERP speaker 1 = (ERP a-sp1 + ERP i-sp1 + ERP u-sp1)/3]. Using *post hoc t* tests, we also analyzed pairwise stimulus differences for speaker and vowel averages at a representative left-temporal electrode.

The timing of oscillatory responses was investigated by computing intertrial phase coherence (ITC) transforms using the EEGLAB toolbox (Delorme and Makeig, 2004). ITC shows the strength (0–1) of phase locking of the EEG signals to stimulus onset. We performed a trial-by-trial time–frequency decomposition using Hanning-windowed sinusoidal wavelets of 1 cycle at 2 Hz, rising linearly to 20 cycles at 70 Hz with a data window length of 556 ms. The resultant ITC transforms reveal intertrial coherence estimates for time/frequency cells centered at (1.0 Hz, 9.7 ms) intervals. Using the same parameter settings, we also computed event-related spectral perturbation (ERSP) estimates. ERSP shows changes in spectral power (in decibels) from the 1000 ms prestimulus baseline.

We first computed average ITC and ERSP measures for the speaker, vowel, and passive control task (see supplemental Fig. 3, available at www.jneurosci.org as supplemental material). To test our neural coding scheme (Fig. 2), we subsequently analyzed the task-dependent phase reorganizations of stimulus-locked oscillatory responses to physically identical stimuli. We analyzed ITC measures in the theta (4–7 Hz), alpha (8–12 Hz), beta (15–25 Hz), and gamma (30–40 Hz) bands, in the time windows of interest as indicated by our ERP results, corresponding to the N1, P2, and P340 intervals, and in the time window preceding the P340 interval (250–310 ms). We computed two measures of ITC from the same set of EEG data elicited by our nine speech conditions: (1) a "vowel grouping" ITC reflecting the average intertrial phase coherence for each of the vowels, regardless of speakers [ITC-vowel = (ITC /a/ + ITC /i/ + ITC /u/)/3], and (2) a "speaker grouping" ITC reflecting the average intertrial phase coherence for each of speakers, regardless of vowels [e.g., ITC-speaker = (ITC sp1 + ITC sp2 + ITC sp3)/3]. Note that both vowel and speaker grouping ITC estimates include all recorded trials, the only difference being the stimulus dimension across which these trials were grouped for calculating estimates of response similarity. Furthermore, as time–frequency decompositions are computed independently for each trial, the overall oscillatory power (ERSP) for speaker and vowel grouping is identical. Speaker and vowel grouping ITCs were analyzed using a repeated-measures ANOVA with task (speaker vs vowel task) and grouping (speaker vs vowel grouping) as within-subjects factors. Task modu-

lation of speaker versus vowel grouping ITCs as well as hemispheric differences in these effects were further analyzed with *post hoc t* tests.

Results

Behavioral responses

Behavioral responses were only required in trials that contained targets (6.5% of trials), corresponding to vowel repetition (vowel task) or speaker repetition (speaker task). Participants correctly performed both the vowel task, 99.5 ± 0.6% (mean ± SD) of all trials correct, and the speaker task, 98.3 ± 2.1% of all trials correct. Trials including targets and/or button presses (correct detections, omissions, and false positives) were not included in the EEG analysis.

Event-related potentials

In each of the three tasks speech stimuli elicited a comparable sequence of typical auditory P1 (~45 ms), N1 (~105 ms), and P2 (~200 ms) responses (see supplemental Fig. 1, available at www.jneurosci.org as supplemental material). In the speaker and vowel, but not in the passive task, these responses were followed by a positive ERP around 340 ms and a sustained activity that lasted until ~800 ms, reflecting the additional cognitive demands posed by both active tasks.

Topographic statistical maps of all ANOVA results are illustrated in supplemental Figure 2 (available at www.jneurosci.org as supplemental material). Because none of the time windows showed a significant task-by-vowel-by-speaker interaction, we focused our further analyses on main effects and two-way interactions, and illustrate ERP effects with vowel (/a/, /i/, /u/) and speaker (sp1, sp2, sp3) averages. Figure 3*a* illustrates ERP effects for vowel and speaker averages at a representative left temporal electrode and topographic statistical maps of the task-by-vowel and task-by-speaker interactions (masked at $p < 0.05$, corresponding to $F_{(2,26)} = 3.4$). The early N1 and P2 windows showed main effects of vowel (see supplemental Fig. 2, available at www.jneurosci.org as supplemental material), which are reflected by the amplitude differences between vowel averages (Fig. 3*a*, left). However, as indicated by the nonsignificant task-by-vowel interactions (see statistical maps in Fig. 3*a*, left), performance of the one-back tasks did not modulate these differences. Note that the P2 window additionally showed a significant vowel-by-speaker interaction, further indicating task-independent processing of physical stimulus differences (supplemental Fig. 2, available at www.jneurosci.org as supplemental material). Conversely, in the P340 window we observed a significant task-by-vowel interaction which was explained by an enhanced differentiation of vowels in the vowel compared with the speaker task.

Similarly, the early N1 and P2 windows showed main effects of speaker (supplemental Fig. 2, available at www.jneurosci.org as supplemental material), indicating task-independent amplitude modulations along the speaker dimension that are reflected by the speaker averages (Fig. 3*a*, right). Only in the P340 window there was a significant task-by-speaker interaction (see statistical maps in Fig. 3*a*, right), which was explained by an enhanced differentiation of speakers in the speaker compared with the vowel task.

The stimulus discrimination matrices in Figure 3*b* further illustrate these effects. For both vowel and speaker averages and for three time windows of interest, we color coded the pairwise statistical differences (*t* values) between responses. It can be seen that in the early windows the discrimination between stimuli is task independent (as shown by the low *t* values in the vowel-speaker difference matrices), whereas in the window centered

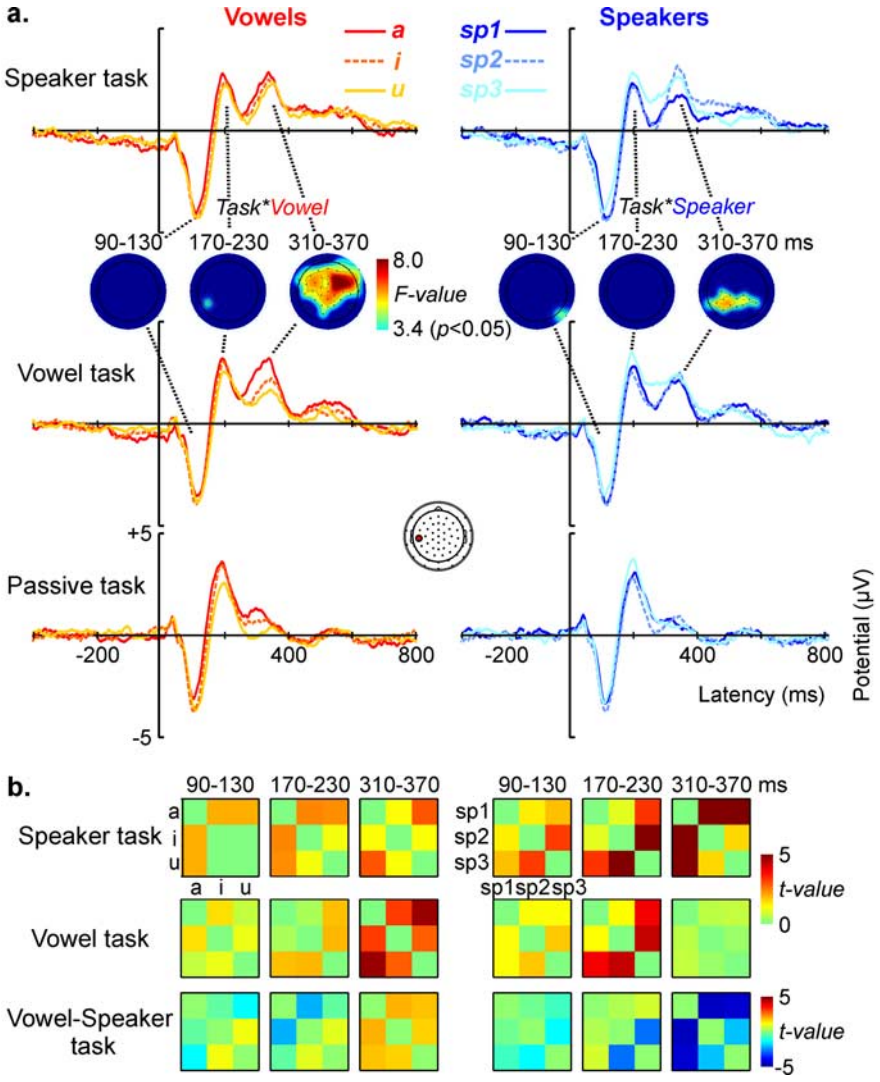

Figure 3. ERP activity elicited by the same speech stimuli grouped across vowels or speakers. ***a***, ERP effects for vowel and speaker averages recorded at a left temporal electrode during the speaker, vowel, and passive control task. Statistical topographic maps show significant task × vowel and task × speaker interactions (thresholded at $p < 0.05$, corresponding to $F_{(2,26)} = 3.4$) in the P340 (310–370 ms), but not in the preceding N1 (90–130 ms) or P2 (170–230 ms) window. In these maps, green-to-red colors encode significance values, the blue color indicates the absence of a significant effect. ***b***, Stimulus discrimination matrices illustrating pairwise statistical differences (color-coded *t* values) between ERP amplitudes of vowels (left) and speakers (right) at the same left temporal electrode. Rows represent discrimination matrices for the speaker, vowel, and vowel − speaker task. In the 310–370 ms window, the difference matrices (vowel − speaker task) indicate increased vowel discrimination during performance of the vowel task (red/orange colors) and increased speaker discrimination during performance of the speaker task (blue colors).

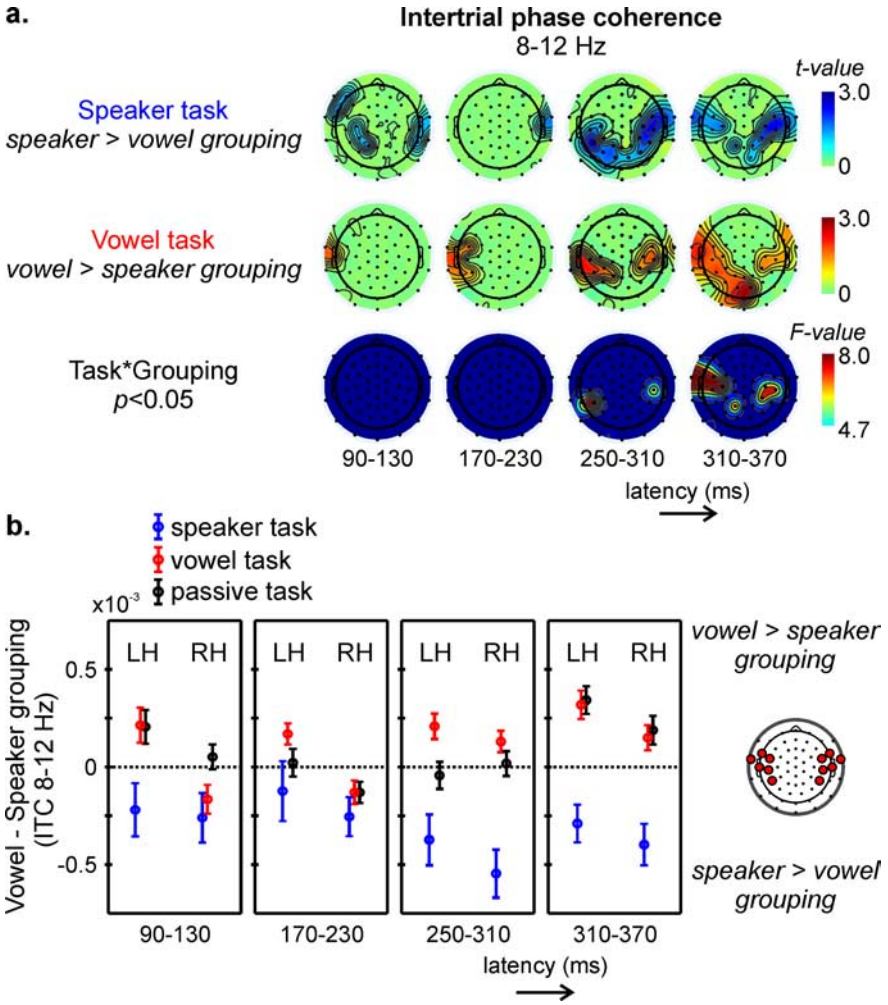

Figure 4. Task-dependent phase reorganization of stimulus-locked alpha oscillations. Intertrial phase coherence (ITC) of alpha activity elicited by all speech stimuli was computed across vowels (vowel grouping) or across speakers (speaker grouping). ***a***, Statistical topographic maps in four different time windows illustrating areas of increased alpha ITC across speakers in the speaker task (top row), across vowels in the vowel task (middle row), and electrodes showing significant task × grouping interactions (bottom row, thresholded at $p < 0.05$, corresponding to $F_{(1,13)} = 4.7$). ***b***, Average (±SE) ITC differences (vowel − speaker grouping) over left (LH) and right (RH) temporal electrodes as measured in the speaker (blue), vowel (red), and passive (black) task in four different time windows.

around 340 ms the discrimination between stimuli is emphasized for the task-relevant stimulus dimension.

Like in both active tasks, the N1 and P2 windows in the passive task showed comparable responses for vowel and speaker averages. In the later P340 window, the amplitudes of vowel and speaker responses paralleled those observed in the vowel task.

In summary, our ERP results indicate task-independent analysis of acoustic and phonetic features of speech stimuli in early time windows (N1–P2), followed by a task-dependent analysis of the behaviorally relevant stimulus dimension in the P340 window. This suggests that during the performance of the one-back vowel and speaker identity tasks, the latter time window is crucial for the maintenance and/or retrieval of abstract representations of the stimuli beyond their acoustic implementation.

Transient phase reorganization of alpha oscillations

In early time windows EEG activity elicited during the speaker, vowel, and passive tasks showed comparable power and phase changes across the different frequency bands (supplemental Fig. 3, available at www.jneurosci.org as supplemental material). Like the grand-average ERP waveforms (supplemental Fig. 1, available at www.jneurosci.org as supplemental material), spectral changes subsequently showed stronger sustained responses after ~250 ms during the speaker and vowel than in the passive task, similarly reflecting the additional cognitive demands posed by both active tasks.

To investigate the role of neuronal oscillations in coding abstract properties of speech stimuli, we focused our further analyses on intertrial phase coherence (ITC) and examined how performance of the vowel versus speaker task influenced the synchronization between EEG responses. Statistical maps in Figure 4*a* represent *t*-maps and time windows of increased intertrial phase synchronization in the alpha band for speakers compared with vowels during the speaker task (top row), for vowels compared with speakers during the vowel task (middle row) and the topographical distribution of significant task-by-grouping interactions (bottom row, masked at $p < 0.05$, corresponding to $F_{(1,13)} = 4.7$). In the time windows corresponding to N1 and P2, these maps indicate a tendency toward task-dependent intertrial phase alignment between speakers (vowels), but the corresponding ANOVAs showed no significant task-by-grouping interaction. The expected pattern of phase alignment along the behaviorally relevant stimulus dimension started in the 250–310 ms time window and most strongly showed in the subsequent P340 window (Fig. 4*a*, bottom row). Thus, at temporal electrodes alpha oscillations demonstrated a significantly increased phase synchronization between speakers compared with vowels during the speaker task (Fig. 4*a*, top row), and the opposite pattern of increased synchronization between vowels compared with speakers in the vowel task (Fig. 4*a*, middle row). None of the other frequency band showed a significant reorganization of responses along the task-relevant stimulus dimension.

The observed alpha ITC effects additionally indicated a right hemispheric bias for the speaker task and a left hemispheric bias for the vowel task. These potential lateralization differences were further examined by comparing average vowel − speaker ITC values over left versus right temporal electrodes (Fig. 4*b*). Statistical tests indicated a significant leftward bias for stimulus-dependent alpha phase reorganization in the P340 window during the vowel task ($p < 0.05$), whereas the right hemispheric bias during the speaker task did not reach significance ($p > 0.1$). None of the other time windows showed significant hemispheric differences. But, like the topographical maps (Fig. 4*b*), the average left and right temporal vowel − speaker ITC values further show a build-up of alpha phase reorganization in the preceding time windows. Finally, similar to our ERP findings, vowel − speaker ITC values confirm a tendency to favor the vowel dimension in the passive task, i.e., increased ITC in the alpha band for

vowel grouping compared with speaker grouping in the P340 window, with a nonsignificant tendency toward a leftward hemispheric bias ($p = 0.09$).

Discussion

We investigated EEG activity evoked by natural speech stimuli while participants performed different tasks that require the tracking of speaker or vowel identity. By examining the task-dependent modulation of ERP and oscillatory responses to acoustically identical stimuli, we could isolate cortical processing specifically related to the extraction and maintenance of speaker-independent vowel information and vowel-independent speaker information.

Time course of stimulus-driven and categorical analysis

The time course of speech processing in the different experimental tasks delineates a stage of largely stimulus-driven analysis followed by task-specific processing of behaviorally relevant stimulus categories. Stimulus-driven analysis is indicated by ERP activity in the early time windows (N1–P2 responses) that is unaffected by our task manipulations. This bottom-up analysis is most likely driven by characteristic acoustic–phonetic features such as fundamental frequency, timbre, or breathiness for speaker discrimination (Murry and Singh, 1980; Klatt and Klatt, 1990; Belin et al., 2004), and the first and second formant frequencies for vowel discrimination (Obleser et al., 2004b; Shestakova et al., 2004). Our findings in these time window are thus consistent with those of previous studies that simultaneously investigated voice and speech sound processing (Obleser et al., 2004a; Shestakova et al., 2004).

Beyond this stage of acoustic–phonetic analysis, we aimed at understanding the task-dependent categorization of physically identical speech stimuli into abstract representations of vowels and speakers. To this end we used experimental tasks that require an active extraction of vowel (speaker) information as well as a suppression of the task-irrelevant stimulus dimension. Furthermore, we used three instead of two (Obleser et al., 2004a) stimulus levels, and demanding one-back tasks instead of simple target detection tasks (Obleser et al., 2004a) or passive listening (Shestakova et al., 2004). In the context of this paradigm, task-dependent processing of speech occurred around 310–370 ms, as indexed by the selective enhancement of P340 amplitude differences for vowels (vowel task) or speakers (speaker task). Interestingly, in this time window, linguistic context effects have been shown to modulate auditory cortical processing of meaningless speech syllables (Bonte et al., 2006). Furthermore, the timing and polarity of the P340 response bear resemblance to those of a voice sensitive response that is elicited during the allocation of attention to voices compared with control sounds (Levy et al., 2003).

The observed amplification of task-relevant stimulus differences (P340) may thus reflect cortical processes underlying the formation and maintenance of abstract vowel/voice representations based on the outcome of a perceptual analysis of acoustic–phonetic stimulus characteristics (N1–P2). Similar processes may be used during the allocation of attention to different classes of auditory objects such as encountered in complex auditory scenes during everyday life (Fritz et al., 2007). Furthermore, the finding that EEG responses in the passive task resemble those of the vowel and not of the speaker task suggests that speech sound and voice information have a different saliency. Thus, without further task demands, the linguistic dimension represents the "default" processing mode for speech. This may also explain why in previous studies EEG/MEG responses to phonemes were shown to reflect acoustic–phonetic features independently of pitch (Poeppel et al., 1997) or speaker (Obleser et al., 2004a; Shestakova et al., 2004), whereas no phoneme-invariant responses to speakers have been reported.

Cortical coding of vowel/speaker invariance

We hypothesized that the analysis of abstract properties of speech (vowel/speaker identity) is encoded by a task-dependent realignment of oscillatory activity to different acoustic events (Fig. 2). Starting at 250 ms, our data indeed show a significantly increased phase alignment of alpha oscillations along the behaviorally relevant stimulus dimension. How could we explain such specific reorganization of alpha phase patterns? Assume that each speech stimulus is characterized by a unique phase pattern of neural responses (Luo and Poeppel, 2007). During initial analysis each of these unique responses is mostly determined by physical stimulus properties and properties of the auditory cortical machinery. Task-dependent selective attention subsequently leads to a top-down modulation of (alpha) oscillations resulting in an alignment of the phases of oscillatory activity evoked by different tokens of either the same vowel (vowel task), or of the same speaker (speaker task). In other words, performance of the vowel task increases the similarity between neural responses to e.g., vowel /a/, independent of who pronounced this vowel, whereas the speaker task increases the similarity between neural responses to e.g., speaker 1, independent of which vowel she/he pronounced. This alpha phase alignment may organize neurophysiological responses such that the behaviorally relevant stimulus dimension is maintained for further processing (von Stein et al., 2000). Interestingly, phase reorganization and ERP amplitude modulations occurred in the same time window. This raises the possibility that alpha phase alignment contributed to the task-dependent enhancement of ERP amplitude differences (Fig. 3), which is consistent with the suggestion that ERP waveforms are modulated by the precise timing of alpha oscillations (Makeig et al., 2002; Klimesch et al., 2007b; Mazaheri and Jensen, 2008).

As both speaker-independent vowel analysis and vowel-independent speaker analysis demonstrated comparable time courses and oscillatory patterning they most likely relied on similar computational mechanisms. The different spatial distribution of the effects, and in particular the left hemispheric bias for the vowel but not the speaker task, indicates the involvement of distinct networks of brain areas. Based on this observation and the putative role of alpha oscillations in modulating functional connectivity between brain areas (von Stein et al., 2000; Kujala et al., 2007; Brancucci et al., 2008), we hypothesize that the observed phase reorganization operates at an interregional level. In particular, this reorganization may mediate temporal binding of distributed neural activity in distinct (auditory) cortical areas subserving the abstract representation of voice and speech (Formisano et al., 2008).

Experimental tasks requiring sound-based analysis of speech typically lead to extensive activation of left hemispheric auditory and higher-order cortical areas (Hickok and Poeppel, 2007). Because the vowel task required phoneme-based analysis, the observed left-hemispheric bias indicates the involvement of a similar network of brain areas. The timing of such task-induced asymmetry may depend on the type of prelexical task used, as previous studies using low-level perceptual tasks reported an earlier leftward lateralization of the N1 response (Poeppel et al., 1996; Vihla and Salmelin, 2003). Furthermore, the frequency band that is modulated may also depend on task demands and the type of experimental stimuli. Thus, oscillations at frequencies

commensurate with the timing of consecutive syllables, especially theta oscillations, have been suggested to track the acoustic structure of intelligible speech (Ahissar et al., 2001; Luo and Poeppel, 2007). Tracking vowel or speaker identity such as investigated in the present study requires abstraction from the acoustic speech signal and relies on retrieval and maintenance of relevant stimulus features and the inhibition of irrelevant features. Our data support and extend previous findings that relate these type of working memory processes to alpha band activity in dedicated cortical areas (Krause et al., 1996; Jensen et al., 2002; Klimesch et al., 2007a). Although alpha oscillations may also show a functional segregation of lower (7–10 Hz) versus higher (10–13.5 Hz) frequencies, with power changes in the former reflecting overall attention effects, and those in the latter reflecting task specific processes (Klimesch et al., 1997), this segregation was not present in the current data. Finally, whereas phase reorganization was not present in the gamma band, such effects may become visible if task demands were changed. Gamma band activity may be more involved when requiring subjects to continuously shift their attention between different speech (vowel/speaker) dimensions (Kaiser et al., 2007), rather than using blocked speaker/vowel tasks, or when presenting ambiguous speech stimuli (Basirat et al., 2008).

Cognitive role of alpha phase synchronization

Most human EEG/MEG studies have focused on estimates of stimulus or task-related power changes, which reflect the amount of underlying synchronous neural activity. Instead, we focused on phase information, which unlike power estimates, directly relates to the timing of neural activity (Engel et al., 2001; Salinas and Sejnowski, 2001). Our findings highlight the importance of the precise timing of the phase of alpha oscillations for cortical information processing. Furthermore, they demonstrate a specific role of alpha phase alignment in the adaptive tuning of neural activity enabling abstract and task-dependent analysis of sensory input. This extends views on the functional significance of alpha band activity in humans which originally linked this frequency band to cortical "idling" (Adrian and Matthews, 1934) and more recently to inhibitory control (Klimesch et al., 2007a). Interestingly, our data provide an empirical demonstration of the view that the precise timing of alpha oscillations "gates" the direction of perceptual (von Stein et al., 2000) or sensory–semantic (Klimesch et al., 2008) processing in the brain. A similarly active and cognitive role is also suggested by recent MEG findings showing task-specific top-down modulations of interareal alpha phase coherence in humans (Kujala et al., 2007). Although the present data specifically focused on sound and voice based analysis of speech, dynamic reorganization of stimulus-evoked oscillatory (alpha) activity may provide a more general neural mechanism underlying auditory object formation. In conclusion, our results show that the transient and goal-dependent realignment of neuronal oscillatory responses enables different abstract representations of the same sensory (auditory) input.

References

Adrian ED, Matthews BH (1934) The Berger rhythm: potential changes from the occipital lobes in man. Brain 57:355–385.

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A 98:13367–13372.

Basirat A, Sato M, Schwartz JL, Kahane P, Lachaux JP (2008) Parieto-frontal gamma band activity during the perceptual emergence of speech forms. Neuroimage 42:404–413.

Beauchemin M, De Beaumont L, Vannasing P, Turcotte A, Arcand C, Belin P, Lassonde M (2006) Electrophysiological markers of voice familiarity. Eur J Neurosci 23:3081–3086.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309–312.

Belin P, Fecteau S, Bédard C (2004) Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8:129–135.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512–528.

Boersma P, Weenink D (2002) Praat 4.0: a system for doing phonetics with the computer [computer software]. Amsterdam: Universiteit van Amsterdam.

Bonte M, Parviainen T, Hytönen K, Salmelin R (2006) Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. Cereb Cortex 16:115–123.

Brancucci A, Penna SD, Babiloni C, Vecchio F, Capotosto P, Rossi D, Franciotti R, Torquati K, Pizzella V, Rossini PM, Romani GL (2008) Neuromagnetic functional coupling during dichotic listening of speech sounds. Hum Brain Mapp 29:253–264.

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

Engel AK, Fries P, Singer W (2001) Dynamic predictions: oscillations and synchrony in top-down processing. Nat Rev Neurosci 2:704–716.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970–973.

Fritz JB, Elhilali M, David SV, Shamma SA (2007) Auditory attention–focusing the searchlight on sound. Curr Opin Neurobiol 17:437–455.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Hopfield JJ, Brody CD (2001) What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration. Proc Natl Acad Sci U S A 98:1282–1287.

Jensen O, Gelfand J, Kounios J, Lisman JE (2002) Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. Cereb Cortex 12:877–882.

Kaiser J, Lennert T, Lutzenberger W (2007) Dynamics of oscillatory activity during auditory decision making. Cereb Cortex 17:2258–2267.

Klatt DH, Klatt LC (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. J Acoust Soc Am 87:820–857.

Klimesch W, Doppelmayr M, Pachinger T, Russegger H (1997) Event-related desynchronization in the alpha band and the processing of semantic information. Brain Res Cogn Brain Res 6:83–94.

Klimesch W, Sauseng P, Hanslmayr S (2007a) EEG alpha oscillations: the inhibition-timing hypothesis. Brain Res Rev 53:63–88.

Klimesch W, Sauseng P, Hanslmayr S, Gruber W, Freunberger R (2007b) Event-related phase reorganization may explain evoked neural dynamics. Neurosci Biobehav Rev 31:1003–1016.

Klimesch W, Freunberger R, Sauseng P, Gruber W (2008) A short review of slow phase synchronization and memory: evidence for control processes in different memory systems?. Brain Res 1235:31–44.

Krause CM, Lang AH, Laine M, Kuusisto M, Pörn B (1996) Event-related EEG desynchronization and synchronization during an auditory memory task. Electroencephalogr Clin Neurophysiol 98:319–326.

Kujala J, Pammer K, Cornelissen P, Roebroeck A, Formisano E, Salmelin R (2007) Phase coupling in a cerebro-cerebellar network at 8–13 Hz during reading. Cereb Cortex 17:1476–1485.

Lee TW, Girolami M, Sejnowski TJ (1999) Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Comput 11:417–441.

Levy DA, Granot R, Bentin S (2003) Neural sensitivity to human voices: ERP evidence of task and attentional influences. Psychophysiology 40:291–305.

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54:1001–1010.

Makeig S, Westerfield M, Jung TP, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ (2002) Dynamic brain sources of visual evoked responses. Science 295:690–694.

Mazaheri A, Jensen O (2008) Asymmetric amplitude modulations of brain oscillations generate slow evoked responses. J Neurosci 28:7781–7787.

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognit Psychol 18:1–86.

Murry T, Singh S (1980) Multidimensional analysis of male and female voices. J Acoust Soc Am 68:1294–1300.

Näätänen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, Allik J, Sinkkonen J, Alho K (1997) Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385:432–434.

Norris D, McQueen JM (2008) Shortlist B: a Bayesian model of continuous speech recognition. Psychol Rev 115:357–395.

Obleser J, Elbert T, Eulitz C (2004a) Attentional influences on functional mapping of speech sounds in human auditory cortex. BMC Neurosci 5:24.

Obleser J, Lahiri A, Eulitz C (2004b) Magnetic brain response mirrors extraction of phonological features from spoken vowels. J Cogn Neurosci 16:31–39.

Poeppel D, Yellin E, Phillips C, Roberts TP, Rowley HA, Wexler K, Marantz A (1996) Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. Brain Res Cogn Brain Res 4:231–242.

Poeppel D, Phillips C, Yellin E, Rowley HA, Roberts TP, Marantz A (1997) Processing of vowels in supratemporal auditory cortex. Neurosci Lett 221:145–148.

Salinas E, Sejnowski TJ (2001) Correlated neuronal activity and the flow of neural information. Nat Rev Neurosci 2:539–550.

Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123:2400–2406.

Shestakova A, Brattico E, Soloviev A, Klucharev V, Huotilainen M (2004) Orderly cortical representation of vowel categories presented by multiple exemplars. Brain Res Cogn Brain Res 21:342–350.

Titova N, Näätänen R (2001) Preattentive voice discrimination by the human brain as indexed by the mismatch negativity. Neurosci Lett 308:63–65.

Vihla M, Salmelin R (2003) Hemispheric balance in processing attended and non-attended vowels and complex tones. Brain Res Cogn Brain Res 16:167–173.

von Stein A, Chiang C, König P (2000) Top-down processing mediated by interareal synchronization. Proc Natl Acad Sci U S A 97:14748–14753.