Brain & Language 157–158 (2016) 63–71

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

The neural dynamics of speech perception: Dissociable networks for processing linguistic content and monitoring speaker turn-taking

Dan Foti [a,*], Felicia Roberts [b]

[a] Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, United States
[b] Brian Lamb School of Communication, Purdue University, 100 N. University Street, West Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 30 June 2015
Revised 27 April 2016
Accepted 1 May 2016
Available online 10 May 2016

Keywords:
EEG
ERP
Speech
Language
Prosody

ABSTRACT

The neural circuitry for speech perception is well-characterized, yet the temporal dynamics therein are largely unknown. This timing information is critical in that spoken language almost always occurs in the context of joint speech (i.e., conversations) where effective communication requires the precise timing of speaker turn-taking—a core aspect of prosody. Here, we used event-related potentials to characterize neural activity elicited by conversation stimuli within a large, unselected adult sample (N = 115). We focused on two stages of speech perception: inter-speaker gaps and speaker responses. We found activation in two known speech perception networks, with functional and neuroanatomical specificity: silence during inter-speaker gaps primarily activated the posterior pathway involving the supramarginal gyrus and premotor cortex, whereas hearing speaker responses primarily activated the anterior pathway involving the superior temporal gyrus. These data provide the first direct evidence that the posterior pathway is uniquely involved in monitoring speaker turn-taking.

1. Introduction

In conversations around the globe, billions of times each day, people shift effortlessly and seamlessly from one speaker to another. This daily exchange of turns at talk, a fundamental structure of human social interaction, remains largely unexplored from a neuroscience perspective. A key entry point for studying these deceptively simple moments is that speaker transition tends to be below the level of awareness, except when something goes awry. The current study leverages that fact to examine neural activity during the processing of delayed and normal turn-taking in conversation. This approach allows for the exploration of the neural pathways involved in processing conversational speech, monitoring the timing of speaker turn-taking, and inferring affect from anomalous speaker transitions.

Over 40 years ago, Sacks, Schegloff, and Jefferson (1974) described several "grossly apparent facts" about conversation and from those observations derived what has become the most widely used, empirically-grounded model of turn-taking for studies of human communication (Roberts & Robinson, 2004). Their model addresses the way in which turns at talk are constructed, ordered, and distributed among speakers. Most important for the current study is their identification of the "transition relevance place," or the moment when one unit of talk comes to an end and a next unit begins (initiated by the same or another speaker). Our goal is to inspect the neural dynamics within this transition space, particularly neural pathways as they relate to the processing of timing cues in turn transitions.

The importance of addressing brain function in more interactional and even true dyadic frameworks, particularly in relation to language and other behavioral processes, has been recently emphasized in neuroscience and psychology (Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012; Scott, McGettigan, & Eisner, 2009; Stephens, Silbert, & Hasson, 2010; Wilson & Wilson, 2005). Indeed, a neuroscience perspective on the motor theory of speech perception suggests that the motor cortex is being recruited primarily to manage the timing of turn-taking in conversation. This is a role that goes beyond deciphering linguistic input, as indicated in some studies (Wilson & Iacoboni, 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004) and implicates the motor cortex in processing social interaction (Scott et al., 2009). Specifically, Scott et al. (2009) propose that speech processing involves two functionally and anatomically distinct neural networks: an anterior pathway that is responsible for decoding word meaning and a posterior pathway that monitors speaker rhythm, rate, and turn-taking. Whereas the anterior pathway encompasses the superior temporal and inferior frontal gyri, it is only the posterior pathway that

* Corresponding author.
E-mail addresses: foti@purdue.edu (D. Foti), froberts@purdue.edu (F. Roberts).

http://dx.doi.org/10.1016/j.bandl.2016.05.001
0093-934X/

encompasses motor areas, along with the supramarginal and inferior frontal gyri. This model is broadly consistent with findings from functional magnetic resonance imaging (fMRI) (Wilson & Iacoboni, 2006; Wilson et al., 2004) and lesion studies (Blank, Bird, Turkheimer, & Wise, 2003; Crinion, Warburton, Lambon-Ralph, Howard, & Wise, 2006), yet direct evidence of functional dissociation between these anterior and posterior pathways is lacking. In the current study we seek to address this empirical gap.

To explore the role of sensorimotor activity in the monitoring of speaker turn-taking, we developed a novel approach for capturing neural dynamics during the processing of naturalistic conversations, using event-related potentials (ERPs). fMRI has excellent spatial resolution but insufficient temporal resolution to precisely map the dynamics of conversation processing. ERPs, meanwhile, have millisecond temporal resolution and are well-suited to characterize how conversations are processed in real time (Bogels, Magyari, & Levinson, 2015; Magyari, Bastiaansen, de Ruiter, & Levinson, 2014). Here, we leveraged the timing information of ERPs to evaluate whether the processing of spoken words and inter-speaker gaps would be linked specifically with the anterior and posterior pathways, respectively—directly testing for the first time the proposed roles of these pathways in speech perception (Scott et al., 2009).

To do this, we examined ERPs elicited by conversation stimuli that were designed with clear turn-taking expectations embedded in their structure: requests and responses. Drawing on the turn-taking model and related descriptive, experimental, and cross-linguistic research, we know that an extended gap where response is expected is a salient social signal of reluctance, uncertainty or other potential cues of "trouble" in the conversation (Brennan & Williams, 1995; Burgoon, Buller, & Guerrero, 1995; Davidson, 1984; Kendrick & Torreira, 2015; Pomerantz, 1984; Roberts & Francis, 2013; Roberts, Francis, & Morgan, 2006; Roberts, Margutti, & Takano, 2011; Swerts & Krahmer, 2005). Building on these findings, we included both normal (200 ms) and extended (700 ms) speaker gaps between mundane requests and affirmative responses, providing a natural window for observing ERPs during two key stages of conversation processing: (a) activity during the silent turn-transition space and (b) activity in response to affirmative responses, following either a normal or extended gap.

We expected that these two stages of conversation processing would elicit unique ERPs: For the inter-speaker gap, the silence between the request and the response, we focused on the auditory stimulus-preceding negativity (SPN), a negative-going slow wave that is involved in stimulus anticipation (Brunia, van Boxtel, & Böcker, 2012) and is maximal at frontal sites (Ohgami, Kotani, Hiraku, Aihara, & Ishii, 2004). For the speaker response, we focused on the P3, which signals the allocation of attention to infrequent or task-relevant stimuli (Polich, 2007). From an affective standpoint, we expected that the affirmative responses following a long gap would be perceived as anomalous and elicit an increased P3 compared to responses following a normal gap. Critically, we also used source estimation to test whether ERPs to the gap and the response would also uniquely map onto the neural pathways as proposed by Scott et al. (2009). We hypothesized that the SPN during the inter-speaker gap would primarily reflect activation within the posterior pathway (i.e., supramarginal gyrus, motor cortex), whereas the P3 to affirmative responses would primarily reflect activation within the anterior pathway (i.e., temporal gyrus).

Given the potential for overlap between activity in the anterior and posterior pathways during conversation processing, we used principal components analysis (PCA) to parse the ERP data. PCA is a highly effective technique for decomposing ERP waveforms into their underlying neural components, separating patterns of activity that overlap in their timing or spatial distributions (Donchin & Heffley, 1979). PCA also improves the accuracy of source localization algorithms compared to localizing the ERP waveform directly (Dien, 2010b); this is important given our aim evaluate the specific activation of the anterior and posterior pathways across stages of processing. To evaluate the likely neural generators at each stage, we used low resolution brain electromagnetic tomography (LORETA) (Pascual-Marqui, 2002, 2007), a localization algorithm that is well-suited to identifying relatively widespread sources of neural activity (i.e., coordinated activity within a network; Pizzagalli, 2007).

2. Material and methods

2.1. Participants

One hundred and thirty-four adults participated in this study. Nineteen were excluded from the current analyses (1 for difficulty hearing, 3 for equipment malfunction, 3 for poor quality EEG data, and 12 for being statistical outliers), leaving 115 in the final sample (Age: $M = 20.23$, $SD = 5.40$; Gender: 67 females, 47 males, 1 declined to answer; Ethnicity: 2 Hispanic, 108 Not Hispanic, 4 declined to answer; Race: 26 Asian, 6 African American, 81 Caucasian, 1 multiracial, 1 declined to answer). Ninety-one participants were current undergraduate students who received course credit for their participation, and the remaining 24 were volunteers from the surrounding community who received monetary compensation for their time. According to a protocol approved by the Institutional Review Board of Purdue University, informed consent was obtained from participants before the experiment.

2.2. Task and materials

2.2.1. Overview of task

A listening task was administered using Presentation software (Neurobehavioral Systems, Inc., Albany, CA) to control the timing of all stimuli. On each trial, participants were presented with an auditory stimulus consisting of a short conversation (10 targets, 4 distractors) that simulated a telephone call between friends. Study participants were thus in the role of a third party overhearing the conversations, as though on a speaker phone. The constructed dialogues concerned mundane themes (e.g., flyers for a school function, going to the gym, going to lunch, homework). After each conversation, study participants provided a judgment, on a six-point scale, about some aspect of the call recipient's affective response to their ostensible friend's request, invitation, or observation about the world. Ratings higher on the scale indicated a perception of more positive affect (e.g., willingness, enthusiasm).

The target dialogues, where the silence manipulations were presented (i.e., 200- or 700-ms inter-speaker gaps), ended with the caller formulating a request (e.g., getting a ride to pick up flyers). The call recipients in these dialogues always answered in the affirmative ("sure") which made compliance with their friend's request lexically specific and clear. Thus, the experimental manipulation (insertion of the silence) modulated the semantics of the "sure" response, coloring them as potentially reluctant or anomalous (Davidson, 1984; Kendrick & Torreira, 2015; Pomerantz, 1984; Roberts & Francis, 2013; Roberts et al., 2006, 2011).

2.2.2. Construction of stimuli

The conversation stimuli were simulated telephone calls based on an actual telephone call between two female friends (Roberts & Robinson, 2004) and were developed using known features of familiarity in telephone interaction (Hopper, 1992; Schegloff, 1979). Each dialogue was approximately 10 s long and began with a friendly greeting sequence followed by the caller reporting some mundane state of affairs (e.g., "I called the copy shop"). After an