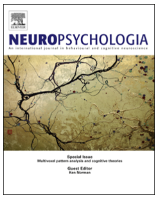
Neuropsychologia 82 (2016) 110–122

Contents lists available at ScienceDirect

Neuropsychologia

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



Cortical processing of phonetic and emotional information in speech: A cross-modal priming study

Erin Diamond [a], Yang Zhang [a,b,c,*]

[a] Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN 55455, USA
[b] Center for Neurobehavioral Development, University of Minnesota, Minneapolis, MN 55455, USA
[c] School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 25 July 2015
Received in revised form
6 January 2016
Accepted 16 January 2016
Available online 18 January 2016

Keywords:
EEG
Speech perception
Emotional prosody
N400
LPR
Cortical oscillation

ABSTRACT

The current study employed behavioral and electrophysiological measures to investigate the timing, localization, and neural oscillation characteristics of cortical activities associated with phonetic and emotional information processing of speech. The experimental design used a cross-modal priming paradigm in which the normal adult participants were presented a visual prime followed by an auditory target. Primes were facial expressions that systematically varied in emotional content (happy or angry) and mouth shape (corresponding to /a/ or /i/ vowels). Targets were spoken words that varied by emotional prosody (happy or angry) and vowel (/a/ or /i/). In both the phonetic and prosodic conditions, participants were asked to judge congruency status of the visual prime and the auditory target. Behavioral results showed a congruency effect for both percent correct and reaction time. Two ERP responses, the N400 and late positive response (LPR), were identified in both conditions. Source localization and inter-trial phase coherence of the N400 and LPR components further revealed different cortical contributions and neural oscillation patterns for selective processing of phonetic and emotional information in speech. The results provide corroborating evidence for the necessity of differentiating brain mechanisms underlying the representation and processing of co-existing linguistic and paralinguistic information in spoken language, which has important implications for theoretical models of speech recognition as well as clinical studies on the neural bases of language and social communication deficits.

1. Introduction

Real-life speech carries both linguistic (e.g. phonetic) and paralinguistic (e.g. emotional prosodic) information. Emotional prosody involves the manipulation of acoustic cues such as fundamental frequency, loudness, and voice quality that allows the speaker to communicate emotion through prosody (Kotz and Paulmann, 2011; Patel et al., 2011). Understanding the brain (cortical as well as subcortical) mechanisms that govern the proper use of the affective cues along with the expression of linguistic content is important for theories on the neural representations of language as well as practical applications such as intervention for individuals with communication difficulties in terms of affective speech comprehension/production (Izdebski, 2008).

A number of important neuroanatomical models have been proposed to explain how the adult human brain represents linguistic messages of the spoken words and sentences including sublexical phonological and lexical semantic details as well as the paralinguistic vocal variations that convey the speaker's emotion and identity. In the dual-stream model for speech perception (Hickok and Poeppel, 2007), the cortical processing system for spoken language features bilateral ventral streams projecting from auditory cortex to middle and inferior posterior temporal regions for mapping sound onto semantic representations and a left-dominant dorsal stream projecting from auditory cortex to parietal-temporal boundary area and frontal regions for mapping sound onto articulatory representations. One limitation of this model is that it did not consider the social-indexical aspects of speech such as emotional prosody and speaker identity. Other models of voice perception endeavor to overcome this limitation by emphasizing functional specialization and heterogeneity of the temporal and frontal cortices, the cortico-subcortical pathways involving regions such as amygdala for emotion processing, and the differential contributions of the right and left hemispheres in vocal emotion processing (Belin et al., 2004; Klasen et al., 2012; Schirmer and Kotz, 2006). For instance, Schirmer and Kotz (2006)

---

[*] Corresponding author at: Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail address: zhanglab@umn.edu (Y. Zhang).

http://dx.doi.org/10.1016/j.neuropsychologia.2016.01.019
0028-3932/

proposes that vocal emotion processing undergoes three stages. The first stage consists of initial bilateral acoustic analyses in the auditory cortex. The second involves projections to specialized voice-selective areas in the superior temporal sulci and gyri for more complex analysis and synthesis of the emotionally salient information. The third continues on with projections to frontal areas for higher-order evaluation and cognition.

The dichotomy of the neuroanatomical models respectively emphasizing either speech or voice processing can be attributed to the different theoretical interests of the representative neuroimaging studies that used entirely different sets of well-controlled linguistic or vocal sounds in contrast to nonspeech stimuli as a control condition. For instance, the linguistic stimuli (syllables, words, or sentences) for a speech perception study were typically spoken in a neutral tone devoid of any emotional significance. While the Schirmer and Kotz (2006) model for voice perception provides both temporal and spatial specifications towards understanding the neural networks for processing emotional prosody, more research is needed to test its predictions, delineate the temporal windows and functional properties (including hemispheric lateralization) of the implicated cortical regions, and understand how the neural networks function and change in relation to emotional valence, sensory modality, age, gender, linguistic experience, and pathological conditions (See Blasi et al. (2011), Izdebski (2008), Wittfoth et al. (2010) for discussion). More importantly, the existing models have not fully addressed how the human brain juxtaposes speech and voice processing in parallel or how it selectively processes one informational dimension in the presence of the other possibly interfering dimension in real time.

In the present study with a visual-auditory priming task, we were particularly interested in exploring cortical mechanisms (i.e., timing, localization, and neural oscillation characteristics) underlying the processing of the emotional aspect of speech prosody as opposed to the recognition of phonetic identity in the same spoken words. Priming paradigms have been employed to evaluate the intersection between pairs of stimuli both within and across modalities. For instance, Greene et al. (2001) explored the effect of within-modal and cross-modal priming on spatio-temporal event processing. They observed an effect for visual priming of auditory targets, but not the reverse. The authors therefore proposed that a visual event provided specific information that could facilitate processing of an auditory target, whereas the auditory primes produced weaker priming effects with the potential to correspond to a wide range of visuals.

In affective priming research, primes and targets are combined to investigate the effects of their congruency or incongruency in a specified aspect of interest. An adjective evaluation task created by Fazio et al. (1986) was an early example. In that experiment, affectively related primes facilitated an evaluative decision of adjective targets, as demonstrated by shorter latencies preceding the adjective evaluation (“good” or “bad”). This has been described as “automatic attitude activation” (for a review, see Fazio (2001)). Affective priming paradigms since then have extended beyond the traditional visual word prime-target pairs to explore interactions between stimulus domains and modalities. Picture primes and written word targets have been used to investigate the neural mechanisms at play during cross-domain visual affective priming paradigms (Zhang et al., 2006; Zhang et al., 2010). Other prime-target pairs have spanned two modalities, combining stimuli such as affective sentence prosody and written words (Schirmer et al., 2002) or musical stimuli and written words (Goerlich et al., 2012). Facial expressions have been used as targets in cross-modal priming paradigms with sentence primes (Czerwon et al., 2013) and musical primes (Lense et al., 2014). Other cross-modal affective priming paradigms have used facial expressions as primes for emotional words (Schirmer and Kotz, 2006) and musical stimuli (Kamiyama et al., 2013).

In order to determine the facilitative effect of primes on targets, a variety of behavioral and brain measures have been utilized in previous studies. For instance, differences in reaction time in response to congruent vs. incongruent pairs are thought to reflect increased or decreased facilitation of target processing. In conjunction with this behavioral measure, event-related potentials (ERPs) are especially valuable as it provides high temporal resolution suitable for investigating brain responses to acoustic and linguistic processing at the millisecond level. Of particular interest to the present study are the N400 and the late positive response (LPR, also referred to as the late positive component [LPC] or late positive potential [LPP]). The N400 component is a negativity occurring approximately 400 ms after the onset of target words involving a violation of meaning (Kutas and Hillyard, 1980). Traditionally, this component has been studied in the context of semantic expectancy violations in sentences (for a review, see Kutas and Federmeier (2011)). But it has been observed in response to incongruency in affective priming paradigms (e.g., Kamiyama et al., 2013; Paulmann and Pell, 2010). A late positive response (LPC) following the N400 has also been observed in affective priming experiments and is generally seen to reflect increased attention to unexpected targets (Werheid et al., 2005; Zhang et al., 2010). The late positivity was identified in various tasks during which the participant was consciously attending to some characteristic of the stimuli, such as congruency (Aguado et al., 2013; Chen et al., 2011; Kamiyama et al., 2013), sound intensity deviation (Chen et al., 2011), or level of arousal (Paulmann et al., 2013). This positive-going deflection is often discussed as a possible variant of the P300 component, reflecting updating working memory (Hajcak et al., 2009; Zhang et al., 2010; also see Donchin and Coles, 1988). The P300 is known to be a neurocognitive index of novelty detection and attentional capture, and its amplitude is strongly dependent on the stimulus context and task demands (e.g., Chennu et al., 2013; Nie et al., 2014).

Localization of the brain processes for the ERP components is another area of investigation. It is generally accepted that language is primarily localized in the left hemisphere whereas emotion processing has greater right hemisphere involvement. However, Schirmer and Kotz (2006) challenge the idea that vocal emotion is uniquely a right hemisphere process. Rather, they propose that prosodic processing is a multi-step process with differential involvement of both hemispheres. Recent investigations into the lateralization of prosody support this model (Iredale et al., 2013; Paulmann et al., 2013; Witteman et al., 2014). In the present ERP study, we attempted to apply source localization method (Luck, 2014; Zhang et al., 2011) to help determine the cortical regions responsible for the generation of possible N400 and LPR responses respectively for phonetic vs. prosodic processing. We expected to see distinct patterns with left-dominant activities for linguistic processing and more complex time-dependent cortical specialization/laterality for emotional processing involving both hemispheres (Zhang, 2011).

In addition to the timing and source localization of neural responses, a relatively new area in ERP research is the application of time-frequency analysis to examine degree of trial-by-trial coherence in cortical rhythms that may give rise to the salient components in the averaged ERP waveforms (Koerner and Zhang, 2015; Luck, 2014). There is a growing body of literature on the cortical rhythms that mediate phonetic and prosodic processing in an audiovisual priming paradigm. The different cortical rhythms are considered to reflect resonant neural networks that code and transfer information across brain regions to support various sensory, motor and cognitive processing. Researchers have investigated how frequency bands are modulated in response to different auditory and visual cues (for a review of EEG coherence, see Weiss and Mueller (2003)). Of particular interest in the current