NEUROSCIENCE
RESEARCH ARTICLE
L. Lu et al. / Neuroscience xxx (2018) xxx–xxx

Semantically Congruent Sounds Facilitate the Decoding of Degraded Images

Lu Lu,[a] Gaoyan Zhang,[a†] Junhai Xu[a†] and Baolin Liu[a,b*]

[a] School of Computer Science and Technology, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin 300350, PR China

[b] State Key Laboratory of Intelligent Technology and Systems, National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, PR China

Abstract—Semantically congruent sounds can facilitate perception of visual objects in the human brain. However, the manner in which semantically congruent sounds affect cognitive processing for degraded visual stimuli remains unclear. We presented participants with naturalistic degraded images and semantically congruent sounds from different conceptual categories in three modalities: degraded visual only, auditory only, and auditory and degraded visual. Functional magnetic resonance imaging was performed to assess variations in brain-activation spatial patterns. In order to account for the facilitation of auditory modulation at different levels, four conceptual categories of stimuli were divided into coarse and fine groups. Conjunction analysis and multivariate pattern analysis were used to investigate integrative properties. Superadditive interactions were found in the visual association cortex and subadditive interactions were observed in the superior temporal sulcus/superior temporal gyrus (STS/STG). Our results demonstrate that the visual association cortex and STS/STG are involved in the integration of auditory and degraded visual information. In addition, the pattern classification results imply that semantically congruent sounds may facilitate identification of degraded images in both coarse and fine groups. Importantly, when naturalistic visual stimuli were further subdivided, facilitation through auditory modulation exhibited category selectivity. © 2018 Published by Elsevier Ltd on behalf of IBRO.

Key words: degraded visual object, multisensory integration, facilitation, category selectivity, multivariate pattern analysis.

INTRODUCTION

To enable effective perception with our multisensory environment, the human brain integrates information from multiple sources into a coherent percept. For example, when watching someone speak, we normally hear the sound of the speech. In such cases, the human brain can effectively integrate information from the visual and auditory modalities via semantically congruent sound. However, the manner in which semantically congruent sounds affect the identification of an obscured visual object is still unclear.

Neurophysiological and functional imaging studies in human and nonhuman primates in the past two decades have advanced our understanding of multisensory integration. The components of a multisensory stimulus are more effectively integrated when they originate from congruent spatial locations (Meredith and Stein, 1986, 1996) and when they occur simultaneously (Miller and D'Esposito, 2005; Senkowski et al., 2007). Two of the simplest forms of multisensory interaction are superadditive and subadditive neural responses. A neuronal response that is larger than the sum of the two responses to the unisensory stimulus is called superadditive. In contrast, responses smaller than the sum of the two responses to the unisensory stimulus, but larger than each response to the unisensory stimulus, are called subadditive (Klemen and Chambers, 2012). Much discussion has centered around the statistical criteria used to classify multisensory integration when comparing bimodal to unimodal conditions using functional magnetic resonance imaging (fMRI) (Beauchamp, 2005; Stein et al., 2009; Love et al., 2011). The three main criteria used in fMRI research are (1) the additive criteria (AV > A + V); (2) the max criteria (AV > max [A, V]); (3) and the mean criteria (AV > mean [A, V]). The max and additive criteria are the most commonly used and discussed metrics for

*Corresponding author at: School of Computer Science and Technology, Tianjin University, Tianjin 300350, PR China. Fax: +86-10-62781789.
E-mail address: liubaolin@tsinghua.edu.cn (B. Liu).
† G. Zhang and J. Xu contributed equally to this work.
Abbreviations: EPI, echo-planar imaging; FDR, false discovery rate; fMRI, functional magnetic resonance imaging; GLM, general linear model; IFG, inferior frontal gyrus; LOBOCV, leave-one-block-out cross-validation; MTG, middle temporal gyrus; MVPA, multivariate pattern analysis; RDM, representational dissimilarity matrix; ROI, region of interest; SNR, signal-to-noise ratio; SPL, sound pressure level; STG, superior temporal gyrus; STS, superior temporal sulcus.

https://doi.org/10.1016/j.neuroscience.2018.01.051

quantifying multisensory integration. Evidence from a cross-modal object recognition study in humans indicates that the posterior superior temporal sulcus and middle temporal gyrus (pSTS/MTG) have enhanced responses when auditory and visual object features are presented together, and that this area is specialized for the integration of different types of information (Beauchamp et al., 2004, 2008). In addition, primary sensory areas also participate in the processing of multisensory interactions (Klemen and Chambers, 2012). For example, cross-modal modulation has been reported to take place in the visual (de Haas et al., 2013) and auditory (Hsieh et al., 2012) cortices.

Recent evidence suggests that the human brain can effectively integrate information from different sensory sources when a semantically congruent stimulus in one sensory modality is presented when another sensory modality is disturbed. A recent behavioral study found that semantically congruent sounds can modulate the identification of masked pictures (Chen and Spence, 2010). Another multisensory speech interactions study has shown that visual speech signals enhance auditory speech comprehension in noisy environments (Ross et al., 2007). An event-related potential study revealed multisensory gains in audio-visual speech recognition at different signal-to-noise ratios (SNRs) when different levels of pink noise were added to speech sounds (Liu et al., 2013). A cross-modal object recognition study reported that superadditive interactions were found for degraded stimuli (the linear interpolation between the original audio-visual stimuli and the random noise phase spectra) in the STS and superior frontal gyrus, and that these interactions successfully modulated audio-visual object categorization (Werner and Noppeney, 2010b). The above study focused on the manner in which auditory and visual stimuli with limited information influence audio-visual integration. However, the environment in which visual objects are identified is often complex. For instance, the visual object may be obscured. In such cases, it remains largely unknown as to where multisensory interactions take place, and what multisensory properties they have when only a visual object is present are corrupted.

In this study, we used naturalistic degraded images and semantically congruent sounds from four conceptual categories to investigate the enhancement of the multisensory integration effect when a visual object is obscured. Participants were presented with audio-visual stimuli in three different modalities: auditory only (A), degraded visual only (V_d), and auditory and degraded visual (AV_d). Conjunction analyses and the classical "max criterion" methods were used to elucidate the regions wherein auditory and degraded visual information were integrated. Furthermore, we investigated whether the facilitation of auditory modulation was characterized by category selectivity by comparing the fine-grained spatial patterns.

EXPERIMENTAL PROCEDURES

Participants

Fourteen participants from Shandong University (mean age, 22 ± 3 years; seven men and seven women) who had normal or corrected-to-normal vision, reported normal hearing, and had no history of neurological or psychiatric illness were enrolled in the fMRI experiment. The study was approved by the local ethics committee. Each participant provided informed consent before the study and received ¥80 after the experiment.

Stimuli

The visual stimuli comprised gray-scale images from four categories: human, animal, mechanical, and nature. These images were downloaded from ImageNet (http://www.image-net.org/). All visual stimuli were presented centrally and were easily distinguished by typical sound characteristics. The size of the each image was edited to 640 × 480 pixels using Adobe Photoshop CS6 (Adobe Systems Incorporated; San Jose, CA, USA). The semantically congruent auditory stimuli, which were selected from the internet, were semantically related to the visual objects. All sound stimuli were edited to have a duration of 2.5 ± 0.5 s (Cool Edit Pro, Syntrillium Software, Adobe Systems Incorporated; San Jose, CA, USA). Auditory stimuli were presented at 80-dB sound pressure level (SPL) (44.1 kHz, 16-bit).

In order to ensure the reliability and objectivity of the stimuli, another 12 participants were recruited to evaluate the stimuli according to familiarity categorization, emotional valence, and semantic consistency (Schneider et al., 2008). All stimuli were presented in an individually randomized order to each participant using E-Prime (E-Studio 2.0, Psychology Software Tools). Immediately after the presentation of each stimulus, the participants were asked to provide responses regarding the following features of the stimuli appearing on the screen:

Familiarity. For the familiarity rating of the stimuli, the participants were instructed to rate the extent to which they were familiar with the object based on a scale ranging from 1 (familiar) to 4 (unfamiliar).

Categorization. The participants allocated each stimulus to one of the four categories (human, animal, mechanical, and nature), which were displayed on the screen.

Emotional valence. For the emotional valence rating of the stimuli, the participants rated the pleasantness of the object represented by the stimulus. The scale ranged from 1 (pleasant) to 5 (unpleasant). A rating of 3 represented neutral valence.

Semantic consistency. For the semantic consistency rating of the stimuli, the participants rated the degree of semantic matching. The scale ranged from 1 (semantic inconsistency) to 4 (semantic consistency).

Stimuli with lower scores on the familiarity and categorization scales, those with biased emotional valence, and those with semantic inconsistency were eliminated. Eight different images and sounds from each category were selected (see Table 1 and Table 2). Gaussian noise (standard deviation = 0.3) was added