NeuroImage 157 (2017) 108–117

Contents lists available at ScienceDirect

NeuroImage

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

Distributed neural signatures of natural audiovisual speech and music in the human auditory cortex

Juha Salmi[a,b,c], Olli-Pekka Koistinen[a], Enrico Glerean[a], Pasi Jylänki[a], Aki Vehtari[a], Iiro P. Jääskeläinen[a], Sasu Mäkelä[a], Lauri Nummenmaa[a,d], Katarina Nummi-Kuisma[e], Ilari Nummi[a], Mikko Sams[a,*]

[a] Department of Neuroscience and Biomedical Engineering (NBE), School of Science, Aalto University, Finland
[b] Advanced Magnetic Imaging (AMI) Centre, School of Science, Aalto University, Finland
[c] Institute of Behavioural Sciences, Division of Cognitive and Neuropsychology, University of Helsinki, Finland
[d] Turku PET Centre, University of Turku, Finland
[e] DocMus Unit, Sibelius Academy, Helsinki, Finland

ARTICLE INFO

Keywords:
Audiovisual
fMRI
Multi-voxel pattern analysis
Music
Speech

ABSTRACT

During a conversation or when listening to music, auditory and visual information are combined automatically into audiovisual objects. However, it is still poorly understood how specific type of visual information shapes neural processing of sounds in lifelike stimulus environments. Here we applied multi-voxel pattern analysis to investigate how naturally matching visual input modulates supratemporal cortex activity during processing of naturalistic acoustic speech, singing and instrumental music. Bayesian logistic regression classifiers with sparsity-promoting priors were trained to predict whether the stimulus was audiovisual or auditory, and whether it contained piano playing, speech, or singing. The predictive performances of the classifiers were tested by leaving one participant at a time for testing and training the model using the remaining 15 participants. The signature patterns associated with unimodal auditory stimuli encompassed distributed locations mostly in the middle and superior temporal gyrus (STG/MTG). A pattern regression analysis, based on a continuous acoustic model, revealed that activity in some of these MTG and STG areas were associated with acoustic features present in speech and music stimuli. Concurrent visual stimulus modulated activity in bilateral MTG (speech), lateral aspect of right anterior STG (singing), and bilateral parietal opercular cortex (piano). Our results suggest that specific supratemporal brain areas are involved in processing complex natural speech, singing, and piano playing, and other brain areas located in anterior (facial speech) and posterior (music-related hand actions) supratemporal cortex are influenced by related visual information. Those anterior and posterior supratemporal areas have been linked to stimulus identification and sensory-motor integration, respectively.

1. Introduction

Our brain integrates auditory and visual information automatically into audiovisual objects. Concordant visual information enhances auditory perception. For instance, viewing concurrent visual speech improves the accuracy of temporal discrimination of the acoustic speech (Vroomen and Stekelenburg 2011). Relatively little is known about audiovisual processing of music, but apparently matching visual information adds to perception of instrumental music, yet in a way that is distinct from audiovisual speech perception (see Saldana and Rosenblum (1993) and Vatakis and Spence (2006)).

1.1. Brain areas involved in speech vs. music

In order to discover the effect of visual stimulation on processing of auditory information in supratemporal auditory cortex, we first have to characterize the areas involved in processing of unimodal auditory stimuli. Music and speech share, for instance, requirement for fine-grained pitch discrimination (Zatorre and Baum, 2012), periodic patterns (Patel, 2003a) and even higher order structures (Patel, 2003b). A few studies have revealed reliable intrahemispheric regional dissociation in cortical processing of complex music and speech features: speech-related spectral irregularity of sounds activates temporal cortex areas, mainly in the middle temporal gyrus (MTG), that are more anterior-lateral to those activated by music-related temporal

* Corresponding author.
E-mail address: Mikko.Sams@aalto.fi (M. Sams).

http://dx.doi.org/10.1016/j.neuroimage.2016.12.005
Received 11 April 2016; Accepted 3 December 2016
Available online 06 December 2016

regularity (Tervaniemi et al., 2006, Santoro et al., 2014). Recent studies utilizing multivariate pattern analysis (MVPA) have detailed different stages in processing unimodal speech and music (Abrams et al., 2011; Norman-Haignere et al., 2015; Rogalsky et al., 2011; Ryali et al., 2010). For instance, Abrams et al. (2011) suggested that unimodal speech and music involve largely the same temporal structure, but distinct spatial patterns to these stimuli can be classified in the inferior frontal gyrus, posterior and anterior superior temporal gyrus (STGp/a) and MTG, and auditory brainstem.

1.2. Brain areas involved in audiovisual modulations

Specific types of concurrent visual input modulate auditory processing in distributed temporal-cortical areas overlapping with those involved in unimodal auditory processing (Kayser et al., 2007). Integration of face and voice (for a review see Campanella and Belin (2007) and Yovel and Belin (2013)), and audiovisual action processing (for a review see Hein and Knight (2008)) are examples of sensory-integration processes that have been widely studied. Visual input modulates activity in multiple areas, including the primary auditory cortex (Sams et al., 1991; Foxe et al., 2002; Pekkola et al., 2005; Kayser et al., 2005) as well as anterior and posterior temporal lobe areas (von Kriegstein et al., 2005; Pekkola et al., 2006; Campanella and Belin, 2007; Perrodin et al., 2014). The role of anterior MTG in coupling the face and voice information, in particular, has been demonstrated in several studies (see Campanella and Belin (2007) and Yovel and Belin (2013)).

Accumulating evidence suggests that audiovisual modulations are largely based on modulation of temporal processing, not changes in the overall response amplitudes (Allman et al., 2008; Iurilli et al., 2012; Lakatos et al., 2007, 2009). For instance, when monkeys are presented with naturalistic sounds accompanied with matching visual stimulus, firing rate of the neurons in the auditory cortex and inter-trial variability of the activation is decreased (Dahl et al., 2010; Kayser et al., 2010).

1.3. Multi-voxel pattern analysis

While the conventional mass-univariate general linear model (GLM) approach is straightforward to implement in studies examining regional activity evoked by isolated stimulus features, it is more problematic when overlapping stimulus features activate distinct multivariate patterns of neural activity within a given region (Ben-Yakov et al., 2012, see also Henson (2006)). Multi-voxel pattern analysis (MVPA) represents an opposite way of modeling, trying to predict stimulus categories using an entire hemodynamic activation pattern, without being restricted to an assumption of certain predefined response function or stimulus model (Norman et al., 2006; Pereira et al., 2008; Mur et al., 2009). By enabling classification of complex stimulus-specific activation patterns even in the absence of regional amplitude changes, MVPA provides a powerful new approach to investigate the mechanisms of audiovisual integration (Pooresmaeili et al., 2014; Gentile et al., 2015; Li et al., 2015; Rohe and Noppeney, 2015). For instance, Li et al. (2015) recently found distributed content-specific (male vs. female, crying vs. laughing) supratemporal activations during audiovisual perception of faces and voices during selective attention to particular features. The effects of matching visual input on processing music and speech, however, remain unclear.

1.4. The aim of the present study

We applied Bayesian logistic regression to classify transient temporal cortex activity patterns measured during audiovisual and auditory speech, singing, and piano playing. The analysis was based on probabilistic classification models that attach a given activation pattern to the most probable one of two or three classes based on linear combinations of the voxel activations, where the signs and absolute values of the voxel coefficients represent the contribution of each voxel to the classification task. By visualizing the posterior probability distributions of the coefficients as brain maps, we expected to reveal neural systems discriminating between audiovisual vs. auditory conditions or between auditory speech, singing and piano playing, likely being represented in complex spatial patterns in distributed neuronal networks (see Abrams et al. (2011), Norman-Haignere et al. (2015), Rogalsky et al. (2011), Ryali et al. (2010) for unimodal studies and Li et al. (2015), Vetter et al. (2014) for audiovisual studies). In order to address this specific research question, we selected a method that is, unlike often used searchlight MVPA approaches (see Mur et al. (2009)), able to detect sparse patterns associated with activity in widely distributed brain networks. To promote sparsity in the posterior solution, the voxel coefficients were given short-tailed Laplace priors, which should improve both the generalizability and interpretability of the solution (Williams, 1995). The performance of the classification models was tested by a cross-validation across 16 participants.

The data were acquired in an fMRI experiment, where participants watched and listened to audiovisual and purely auditory versions of songs that were either spoken, sung, or played with a piano. The visual input in the speech and singing conditions was the face of the speaker/singer, and in the piano conditions participant saw the players finger movements on a keyboard. Singing condition that contained the acoustic structure of music, but had the same voice and mostly similar visual information as in speech condition, was expected to provide additional information about the effects of the visual input type (facial processing in singing vs. hand action in piano playing) and specific spectrotemporal characteristics of music (tone vs. voice) on auditory processing. By using spoken lyrics of the songs in the speech condition we were able to control for semantic and syntactic structures, as well as tempo. The trade-off was that the stimulus was not the most common type of narrative speech but more like listening to poetry reading. As many previous studies (Beauchamp et al., 2004; Romanski and Hwang, 2012; Wayne and Johnsrude, 2012; Conrad et al., 2013; Li et al., 2015), we used complex naturalistic stimuli in order to activate widespread temporal cortex areas associated with audiovisual processing. Such complex stimulation is important also, because it includes nuanced spectro-temporal features that are critical in discriminating between real-life music and speech. Half of the trials contained synchronous matching auditory and visual stimuli, and the other half only auditory stimuli that were identical to those in audiovisual stimuli. Identical auditory stimuli thus canceled acoustic differences related to differences between audiovisual vs. auditory speech, singing, and piano conditions.

We had two predictions: 1) Coherent visual input mostly amplifies processing within the set of brain areas dedicated to processing auditorily presented speech and music, 2) or there are distinct brain areas that specifically contribute to multimodal integration, not involved in auditory processing per se. Furthermore, we expected that visual stimulation containing facial movements would modulate the activity in anterior MTG (Campanella and Belin, 2007; Yovel and Belin, 2013), and that viewing visual hand actions (piano) would, in turn, modulate the activity in the dorsal auditory stream involved in spatial processing and sensorimotor integration (Rauschecker, 2011).

2. Materials & methods

2.1. Participants

We studied 16 healthy participants (6 females; 1 left handed; age range 21–40 years, M_{age} = 28 years, SD_{age} = 2.6 years) with no neurological or psychiatric illnesses or contraindications for functional magnetic resonance imaging, and with normal vision and hearing. All were native Finnish speakers. Seven participants reported music as their hobby, five had experience in playing a musical instrument, and three had studied music theory (15, 10, and 3 years). The study was approved by the Ethical Committee of Hospital District of Helsinki and