Neuropsychologia 75 (2015) 402–410

Contents lists available at ScienceDirect

Neuropsychologia

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

The contribution of dynamic visual cues to audiovisual speech perception

Philip Jaekl [a,*], Ana Pesquita [b], Agnes Alsius [c], Kevin Munhall [c], Salvador Soto-Faraco [d,e]

[a] Center for Visual Science and Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA
[b] UBC Vision Lab, Department of Psychology, University of British Colombia, Vancouver, BC, Canada
[c] Department of Psychology, Queen's University, Kingston, ON, Canada
[d] Centre for Brain and Cognition, Department of Information Technology and Communications, Universitat Pompeu Fabra, Spain
[e] Institució Catalana de Recerca i Estudis Avançats (ICREA), Spain

ARTICLE INFO

Article history:
Received 18 September 2014
Received in revised form
11 June 2015
Accepted 18 June 2015
Available online 20 June 2015

Keywords:
Speech-in-noise
Visual form
Visual motion
Visual pathways
Biological motion
Configural
Audiovisual enhancement

ABSTRACT

Seeing a speaker's facial gestures can significantly improve speech comprehension, especially in noisy environments. However, the nature of the visual information from the speaker's facial movements that is relevant for this enhancement is still unclear. Like auditory speech signals, visual speech signals unfold over time and contain both dynamic configural information and luminance-defined local motion cues; two information sources that are thought to engage anatomically and functionally separate visual systems. Whereas, some past studies have highlighted the importance of local, luminance-defined motion cues in audiovisual speech perception, the contribution of dynamic configural information signalling changes in form over time has not yet been assessed. We therefore attempted to single out the contribution of dynamic configural information to audiovisual speech processing. To this aim, we measured word identification performance in noise using unimodal auditory stimuli, and with audiovisual stimuli. In the audiovisual condition, speaking faces were presented as point light displays achieved via motion capture of the original talker. Point light displays could be isoluminant, to minimise the contribution of effective luminance-defined local motion information, or with added luminance contrast, allowing the combined effect of dynamic configural cues and local motion cues. Audiovisual enhancement was found in both the isoluminant and contrast-based luminance conditions compared to an auditory-only condition, demonstrating, for the first time the specific contribution of dynamic configural cues to audiovisual speech improvement. These findings imply that globally processed changes in a speaker's facial shape contribute significantly towards the perception of articulatory gestures and the analysis of audiovisual speech.

1. Introduction

Speech perception is an inherently multisensory process. Auditory and visual speech signals are both produced by a speaker's vocal apparatus and consequently their time-varying phonetic and kinematic patterns are strongly correlated over time (Chandrasekaran et al., 2009; Girin et al., 2001; Rosenblum, 2008; Summerfield, 1992; Yehia et al., 1998). Humans capitalise on these audiovisual correlations during face-to-face speech perception to decode the message more effectively. For example, in noisy environments, seeing a speaker's facial articulations can dramatically improve speech detection, identification and comprehension (Erber, 1975; Girin et al., 2001; Grant and Seitz, 2000; Grant and Walden, 1996; Macleod and Summerfield, 2009; Robert-Ribes et al., 1998; Ross et al., 2007; Shahin and Miller, 2009). Remarkably, correlated visual information from the speaker's facial movements improves detection of auditory speech in noise even when neither visual nor auditory speech segments can be identified reliably in isolation (e.g., Rosen et al., 1981). Such findings underscore the presence of strong redundancies between visual and auditory properties of speech and the brain's sensitivity to this crossmodal correspondence (Nath and Beauchamp, 2011; Rosenblum, 2008). It is clear that these audio–visual speech correlations unfold over time, given the temporal nature of the speech signal itself. Therefore, visual speech is a complex biological signal spanning across different levels of analysis (Campbell, 2008), from edge detection to the processing of visual prosody. Here we address the nature of visual information contained in the speech signal that is relevant for audiovisual interactions in speech perception. We based our approach in the possible distinct

* Corresponding author.
E-mail address: pjaekl@cvs.rochester.edu (P. Jaekl).

http://dx.doi.org/10.1016/j.neuropsychologia.2015.06.025
0028-3932/

contribution of configural versus local motion cues in the decoding of visual speech.

1.1. Luminance-defined motion and dynamic configural information

The importance of configural cues in vision was famously demonstrated in 1980 with the Thatcher effect (Thompson, 1980). It was shown that gross deviations in facial features that immediately pop out in an edited photograph, could not be detected when the same picture was seen upside down. Configural cues consist of the relative arrangement between features (e.g. the metric distance between lower and upper lip; Collishaw and Hole, 2000; Rhodes et al., 1993; Sergent, 1984) and are derived from the structure of the face and mouth, extracted from the spatial relation between the key facial features (e.g., the distance between the lips and the chin). As the face and mouth change dynamically, the visual articulatory pattern is revealed over time, thus providing *dynamic configural* information (Giese and Poggio, 2003; Lange et al., 2006).

These configural cues are based on samples of the instantaneous configuration of a group of points or features – 'global form' – within a dynamic display. Importantly, temporal integration of a sequence of configurations can reveal structural information regarding stereotyped or common patterns of image changes – sequences that occur, for example, while viewing a walking or talking human figure. The hypothesis that such sequences can reveal human dynamic information has been directly tested using point-light walking displays. For example human data was well modelled by Lange et al. (2006), using a template-matching approach. Here, human configurations of walking stimuli were sampled over time by the model and these sample sequences matched to temporally extended templates of configuration sequences stored in memory. Here, we investigate the role of configural information in the context of audiovisual speech. Specifically, we investigate the possibility that learned sequences of configural changes in visual speech may interact with auditory information and improve word identification of speech in noise.

As mentioned above, the visual speech signal also carries basic motion information, signalled primarily by luminance contrast information (e.g. light spots on a dark background). Here, luminance-defined motion information is initially not extracted directly from global or holistic processing of a set of features but, instead, it is primarily conveyed by local spatiotemporal changes in luminance (see Lu and Sperling (2001) for a comprehensive review). These local (featural) luminance-defined changes are processed by visual mechanisms sensitive to edges or contours defined by differences in luminance intensity.

Luminance-defined local motion information may subsequently be integrated to form global *motion* patterns that are composed of some or many local motions and may reflect, for example, the opposing motion of the legs of a walking person or, the upper and lower lips of a talker (represented in Fig. 1b by a set of arrows pointing in several motion directions). Configural changes and luminance-defined local motion signals are intrinsically related and normally change isomorphically but do not necessarily rely on the same processing mechanisms for perception (Goodale and Milner, 1992; Livingstone and Hubel, 1988). The critical distinction is that luminance-defined local motion is defined by its strong reliance on local spatiotemporal changes in luminance and the processing mechanisms specialized for these cues, whereas dynamic configural cues may be derived from changing form information conveyed through other visual properties in addition to luminance, such as colour (Garcia and Grossman, 2008; Lu and Sperling, 2001). The mechanisms for processing dynamic configural information do not entirely overlap with those required for luminance-defined motion processing.

Notably, configural and luminance-motion cues engage dissociable brain networks, roughly corresponding respectively to the well-known ventral (configuration/form) and dorsal (motion) divisions of the visual system (Goodale and Milner, 1992; Hubel and Livingstone, 1987; Livingstone and Hubel, 1988). The contribution/dissociation of dynamic configural cues and luminance-defined local motion cues towards biological motion perception have been examined extensively (Beintema and Lappe, 2002; Beintema et al., 2006a; Casile and Giese, 2005; Garcia and Grossman, 2008; Giese and Poggio, 2003; Lange et al., 2006; Thompson et al., 2005). Neurophysiologically plausible explanations of biological motion perception have used both dynamic configural and luminance-defined local motion information, as input signals conferring biological motion perception (Giese and Poggio, 2003; Thirkettle et al., 2009a), emphasising the utility of either cue for the perception of walking stimuli.

Primate studies have revealed neurons that respond selectively to static configurations of human forms in the superior temporal sulcus (Vangeneugden et al., 2011, 2009), in the fusiform area (Michels et al., 2005; Peelen and Downing, 2005) and in occipital areas (Grossman and Blake, 2002; Michels et al., 2005; Peelen and Downing, 2005). Configurally based approaches to biological motion – motion of a body or its parts – analysis have modelled responses that are sequence-selective for specific global configural change patterns that unfold over time (Giese and Poggio, 2003; Lange and Lappe, 2006). Alternatively, luminance-defined local motion approaches to understanding biological walking motion perception may be modelled using global patterns extracted from

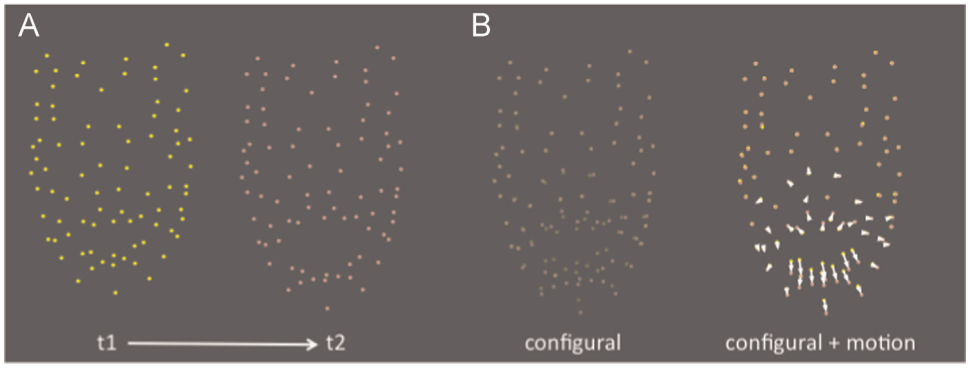

Fig. 1. (A) Point-light facial configuration during a speech utterance at two points in time (t1 and t2; stimuli in experimental display were always yellow). (B) Left panel: Time-varying spatial information can be conveyed by configural changes for isoluminant stimuli. The configural change from t1 to t2 is shown by the images in (A) superimposed, depicting a mouth-opening gesture. Right panel represents the combination of configural and local luminance-motion changes in a higher contrast display. Arrows show local motion paths from t1 to t2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)