Available online at www.sciencedirect.com

ScienceDirect

Vision Research 47 (2007) 1075–1093

Vision Research

www.elsevier.com/locate/visres

Feature-based processing of audio-visual synchrony perception revealed by random pulse trains

Waka Fujisaki [1], Shin'ya Nishida *

NTT Communication Science Laboratories, NTT Corporation, Atsugi, Kanagawa 243-0198, Japan

Received 22 June 2006; received in revised form 6 January 2007

Abstract

Computationally, audio-visual temporal synchrony detection is analogous to visual motion detection in the sense that both solve the correspondence problem. We examined whether audio-visual synchrony detection is mediated by a mechanism similar to low-level motion sensors, by one similar to a higher-level feature matching process, or by both types of mechanisms as in the case of visual motion detection. We found that audio-visual synchrony–asynchrony discrimination for temporally dense random pulse trains was difficult, whereas motion detection is known to be easy for spatially dense random dot patterns (random dot kinematograms) due to the operation of low-level motion sensors. Subsequent experiments further indicated that the temporal limiting factor of audio-visual synchrony discrimination is the temporal density of salient features not the temporal frequency of the stimulus, nor the physical density of the stimulus. These results suggest that audio-visual synchrony perception is based solely on a salient feature matching mechanism similar to that proposed for high-level visual motion detection.

Keywords: Audio-visual; Temporal synchrony; Correspondence problem; Temporal crowding; Saliency based matching

1. General introduction

In our daily lives, we encounter environments where visual signals are often accompanied by concomitant auditory signals arising from the same event. Human observers integrate such an audio-visual signal pair into a coherent percept of a single multi-modal event. Since it is unlikely that audio-visual signals of the same physical cause are far separated in time, it is not surprising that physical temporal proximity (approximate synchrony or simultaneity) is a critical condition for subjective audio-visual integration (Munhall, Gribble, Sacco, & Ward, 1996; Shams, Kamitani, & Shimojo, 2002; Watanabe & Shimojo, 2001). However, previous studies have not fully revealed how the human sensory system detects audio-visual synchrony. Specifically, it remains an open problem as to which level of sensory processing is involved and what sort of representations and algorithms are used for temporal matching (Marr, 1982).

Several lines of evidence argue against a simple view that sensory modalities are separate modules that interact with each other only at post-sensory processing levels (Shimojo & Shams, 2001; Spence & Driver, 2004). Neurophysiological studies have shown the existence of multisensory neurons in the superior colliculus and polisensory cortex, as well as the existence of cross-modal interactions even in primary sensory areas (Schroeder & Foxe, 2005; Stein & Meredith, 1993). It has also been shown that early components of event-related potentials could be influenced by redundant audio-visual information (Lebib, Papo, de Bode, & Baudonniere, 2003; Musacchia, Sams, Nicol, & Kraus, 2006; van Wassenhove, Grant, & Poeppel, 2005). Behaviorally, it has been suggested that the ventriloquist effect, an illusory visual capture of the spatial location of an auditory signal occurs at early pre-attentive levels, since it does not depend on the direction of automatic or

* Corresponding author. Fax: +81 46 240 4716.
E-mail address: nishida@brl.ntt.co.jp (S. Nishida).

[1] Research Fellow of the Japan Society for the Promotion of Science.

0042-6989/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2007.01.021

deliberate visual attention (Bertelson, Vroomen, de Gelder, & Driver, 2000; Vroomen, Bertelson, & de Gelder, 2001), and can modulate the location of auditory attention (Spence & Driver, 2000). Other phenomena that could be interpreted as suggesting early binding of audio-visual signals include the enhanced audibility/visibility of coupled audio-visual signals (Odgaard, Arieh, & Marks, 2004; Sheth & Shimojo, 2004; Stein, London, Wilkinson, & Price, 1996),[2] perceptual integration of visual and auditory motion signals (Meyer, Wuerger, Rohrbein, & Zetzsche, 2005; Soto-Faraco, Spence, & Kingstone, 2005),[3] visual modulation of auditory perception (McGurk & MacDonald, 1976; Soto-Faraco, Navarra, & Alsius, 2004), and auditory modulation of visual perception (Gebhard & Mowbray, 1959; Recanzone, 2003; Sekuler, Sekuler, & Lau, 1997; Shimojo & Shams, 2001; Shipley, 1964). It is possible to interpret these findings (audio-visual interactions in anatomically peripheral brain areas, temporally fast responses, or preattentive sensory processes) to indicate that at least some audio-visual interactions reside at relatively early processing levels.

However, any argument about the level of processing is likely to raise controversy unless there is a conceptual clarification of the potential mechanisms for each level. In examining the level of processing for audio-visual synchrony detection, our psychophysical study was intended to investigate functional levels, which may or may not correspond to anatomical hierarchies. As a conceptual framework, we conceived a concrete hypothesis about potential low- and high-level functional mechanisms for audio-visual synchrony detection by referring to the mechanisms of a similar, and more extensively studied problem — visual motion detection. Computationally, visual motion detection is analogous to audio-visual temporal synchrony detection in the sense that both solve the correspondence problem (Marr, 1982). That is, while the task of audio-visual synchrony detection is to find correspondence between signals from different modalities on the basis of temporal proximity, the task of visual motion detection is to find correspondence between visual signals on the basis of spatiotemporal proximity (Dawson, 1991; Ullman, 1979). Although the same problem is shared with other perceptual processes including binocular stereopsis and binaural sound localization (Banks, Burr, & Morrone, 2006), a merit of comparison with motion detection is that we have good models for low- and high-level motion processing.[4]

The extensive study of visual motion detection has so far revealed the existence of at least two types of detection mechanisms (e.g., Braddick, 1974; Cavanagh & Mather, 1989; Chubb & Sperling, 1988; Lu & Sperling, 2001; Nishida & Ashida, 2001; Nishida, Ledgeway, & Edwards, 1997; Nishida & Sato, 1992; Nishida & Sato, 1995). One exploits low-level specialized sensors that compute motion directly from raw sensory signals. Braddick (1974) introduced the notion of low-level motion sensors under the name of the short-range process to account for his finding that a random dot kinematogram is correctly perceived only with short displacements. Nowadays, this low level motion detecting mechanism is more often called the first-order motion sensor, since later studies showed that it is not characterized by the operating spatial range, but by the type of input signals (a first-order spatial property, luminance) (Cavanagh & Mather, 1989; Chang & Julesz, 1983; Chubb & Sperling, 1988). The computation of this mechanism is considered to be a cross-correlation of spatiotemporally separate luminance signals (Reichardt, 1961) with peripheral spatiotemporal bandpass filters (van Santen & Sperling, 1985), or nearly mathematically equivalent computation of spatially local motion energy within a given band of spatiotemporal frequency (Adelson & Bergen, 1985; Watson & Ahumada, 1985). The use of raw sensory signals by this mechanism is suggested by the finding that the motion sensors are most sensitive within the whole visual system when the stimuli are low-spatial-frequency and high-temporal-frequency luminance modulations (Watson & Ahumada, 1985; Watson & Robson, 1981). The visual system may also include low-level motion sensors specialized for detecting movements of second-order spatial or temporal properties, such as contrast modulation and flicker modulation (Cavanagh & Mather, 1989; Chubb & Sperling, 1988; Lu & Sperling, 1995b; Nishida et al., 1997). These second-order motion sensors are suggested to have a structure similar to the first-order motion sensor except for non-linear preprocessing (Chubb & Sperling, 1988).

In addition to these low-level motion sensors, the visual system has a high-level motion mechanism, which has been called the long-range motion process (Braddick, 1974), attentive tracking (Cavanagh, 1991, 1992), or the third-order motion mechanism (Lu & Sperling, 1995a, 1995b, 2001). The existence of this mechanism was inferred from motion perceptions that cannot be detected by first-order motion sensors, or by second-order motion sensors (Cavanagh, 1991; Lu & Sperling, 1995a). A representative stimulus is the inter-attribute apparent motion, in which the first element distinguished from the background in an arbitrary stimulus dimension (e.g., luminance, color, texture, depth, motion) is perceived to move to the second element defined by another dimension (Cavanagh, Arguin, & von Grünau, 1989; Lu & Sperling, 1995a). Lu and Sperling (1995a, 1995b, 2001) propose that this high-level motion computation uses feature-independent, common representation, which they called stimulus "salience", as input of a motion detector (a spatiotemporal comparator similar to those for low-level motion sensors). They used the term salience to describe the assumed neural process that underlies the

[2] Some effects however might be explained by a response bias change (Odgaard, Arieh, & Marks, 2003).

[3] Audio-visual perceptual integration is not always supported (Alais & Burr, 2004).

[4] Binocular stereopsis is also known to involve multiple mechanisms (Julesz, 1971; Liu, Stevenson, & Schor, 1994; Ramachandran, Rao, & Vidyasagar, 1973; Wilcox & Hess, 1997), but it is open as to whether it includes a high-level feature matching mechanism as proposed for motion processing (Cavanagh, 1991, 1992; Lu & Sperling, 1995a, 1995b, 2001).