Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 57 (2007) 396–414
Journal of Memory and Language
www.elsevier.com/locate/jml

The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception

Emiel Krahmer *, Marc Swerts

Communication and Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Received 9 January 2006; revision received 1 May 2007
Available online 24 July 2007

Abstract

Speakers employ acoustic cues (pitch accents) to indicate that a word is important, but may also use visual cues (beat gestures, head nods, eyebrow movements) for this purpose. Even though these acoustic and visual cues are related, the exact nature of this relationship is far from well understood. We investigate whether producing a visual beat leads to changes in how acoustic prominence is realized in speech, and whether it leads to changes in how prominence is perceived by observers. For Experiment I (''making beats'') we use an original experimental paradigm in which speakers are instructed to realize a target sentence with different distributions of acoustic and visual cues for prominence. Acoustic analyses reveal that the production of a visual beat indeed has an effect on the acoustic realization of the co-occurring speech, in particular on duration and the higher formants (F_2 and F_3), independent of the kind of visual beat and of the presence and position of pitch accents. In Experiment II (''hearing beats''), it is found that visual beats have a significant effect on the perceived prominence of the target words. When a speaker produces a beat gesture, an eyebrow movement or a head nod, the accompanying word is produced with relatively more spoken emphasis. In Experiment III (''seeing beats''), finally, it is found that when participants *see* a speaker realize a visual beat on a word, they perceive it as more prominent than when they do not see the beat gesture.

Keywords: Gestures; Beats; Facial expressions; Audio-visual speech; Acoustics; Speech production; Speech perception

Introduction

Speakers have a large repertoire of potential cues at their disposal which they may use to support what they are saying, including gestures and facial expressions. There is a growing awareness that spoken language and manual gestures are closely intertwined (e.g., Goldin-Meadow, 2003; Mayberry & Nicoladis, 2000; Wagner, Nusbaum, & Goldin-Meadow, 2004), as are spoken language and facial expressions or head movements (e.g., Barkhuysen, Krahmer, & Swerts, 2005; Krahmer & Swerts, 2005; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Srinivasan & Massaro, 2003; Swerts & Krahmer, 2005). Still, the exact relation between auditory speech and visual gestures (of face, arm and body) is far from well understood. In this paper, we take a closer look at a particular kind of gesture that has received relatively little attention so far, namely *beats*. We are interested in the effects of these beat gestures on *prominence*, that is, the relative accentual strength with which words are realized in a spoken utterance. More specifically, we look at whether produc-

* Corresponding author. Fax: +31 13 4663110.
E-mail address: E.J.Krahmer@uvt.nl (E. Krahmer).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.06.005

ing a visual beat leads to a change in how prominence is realized in speech, and whether it leads to a change in how prominence is perceived by observers.

That speech and gesture are related is an old observation (McNeill, 1992 refers to Quintilian's *Institutio Oratoria* from 93 AD as an early source), and one that has been made in various disciplines. In work on the origin of speech, for instance, various researchers have suggested that language may originally have been encoded in gestures rather than in speech (e.g., Corballis, 1992; Fitch, 2000; Holden, 2004). This suggestion is based on the claim that the same brain areas control manual gestures and articulatory gestures, and it has indeed been proposed that a single mechanism may account for the underlying control of both manual gestures and oral gestures required for speech (e.g., Flanagan, Feldman, & Ostry, 1990). According to Holden (2004), evolutionary changes in the brain areas that control gestures might be responsible for the development of our language capacity.

In studies of speech perception, to give a second example, gestures have also played an important role. One of the central questions in speech perception is how listeners are able to map acoustic signals to linguistic elements such as phonemes. Three main theoretical perspectives on this issue have been developed in the past 50 years (Diehl, Lotto, & Holt, 2004). Two of these are based on the assumption that listeners recognize speakers' articulatory gestures, such as lip or tongue movements; intended gestures in *motor theory* (e.g., Liberman, 1957; Liberman & Mattingly, 1985) and real gestures in the *direct realist theory* (e.g., Fowler, 1991, 1996). Both these theories have claimed that the fact that human listeners use visual as well as acoustic information in speech perception (e.g., Dodd & Campbell, 1987; Schwartz, Berthommier, & Savariaux, 2004; Tuomainen, Andersen, Tiippana, & Sams, 2005) offers support for a gestural account of speech perception. A prime example of this is the *McGurk effect* (McGurk & MacDonald, 1976) in which an auditory /ba/ combined with a visual /ga/ is perceived as /da/ by most people. Interestingly, the McGurk effect not only works when articulatory gestures are *seen*; Fowler and Dekle (1991) had listeners identify synthetic /ba/ and /ga/ stimuli, while simultaneously touching the mouth of a talker producing either /ba/ or /ga/. Participants could not see the speaker, but still this haptic variant of the McGurk effect gave rise to reliable evidence of cross-modal effects on perception.

More recently, detailed analyses of speakers have confirmed that they produce speech and manual gestures in tandem, and among researchers in this field there appears to be a general consensus that speech and manual gesture should be seen as two aspects of a single process (e.g., Kendon, 1980, 1997; McNeill, 1992). But the jury is still out on *how* speakers co-produce speech and manual gestures. This can be illustrated by comparing various models for the combination of spoken language and manual gestures that were recently proposed, such as those of Kita and Özyürek (2003), Krauss, Chen, and Chawla (1996), and de Ruiter (2000), all based on the speech production model described by Levelt (1989). What these proposals have in common is the addition of a new gesture stream, which has a shared source with the speech production module but is otherwise essentially independent from it. The main difference between the proposed models lies in the location where the two streams (speech and gesture) part. According to Krauss and co-workers, for instance, this happens before conceptualization, while both de Ruiter as well as Kita and Özyürek argue that the separation takes place in the conceptualizer. McNeill and Duncan (2000) take a markedly different view and argue that speech and gesture should not be delegated to different streams, but rather are produced in close connection with each other, based on what they call "growth points". Thus, even though these researchers agree that speech and manual gestures are closely related, they disagree on how tight this relation is.

It is conceivable that different *kinds* of gestures should be integrated in different ways in speech models, although this aspect of speech–gesture interaction is still largely unexplored. In the literature on manual gestures, a distinction is usually made between representational gestures, "gestures that represent some aspect of the content of speech" (Alibali, Heath, & Myers, 2001) and beat gestures that do not represent speech content (see e.g., Alibali et al., 2001; Krauss et al., 1996; McNeill, 1992). Most of the proposed models focus on representational gestures, such as gestures indicating shape ("round") or motion ("upwards"). In fact, Alibali et al. (2001:84) stress "the need for further study of beat gestures and their role in speech production and communication."

A typical beat gesture is a short and quick flick of the hand in one dimension, for example up and down, or back and forth (McNeill, 1992). These gestures look somewhat like the gestures a conductor makes when beating music time (hence their name); they are sometimes also called "batons" (Efron, 1941; Ekman & Friesen, 1969), in reference to the slender rod used by conductors to direct an orchestra. There is an ongoing, general discussion about what, if anything, different kinds of gestures communicate (e.g., Goldin-Meadow & Wagner, 2005). According to Alibali et al. (2001) beat gestures have no semantic content. Still that does not mean that beats are without communicative value. According to McNeill (1992:15), "the semiotic value of a beat lies in the fact that it indexes the word or phrase it accompanies as being significant (...) for its discourse pragmatic content." A beat thus provides extra prominence for a word, for instance, because it expresses new information (McNeill 1992:169–170).