Guest Editorial

Gesture and speech in interaction: An overview

Abstract

Gestures and speech interact. They are linked in language production and perception, with their interaction contributing to felicitous communication. The multifaceted nature of these interactions has attracted considerable attention from the speech and gesture community. This article provides an overview of our current understanding of manual and head gesture form and function, of the principle functional interactions between gesture and speech aiding communication, transporting meaning and producing speech. Furthermore, we present an overview of research on temporal speech-gesture synchrony, including the special role of prosody in speech-gesture alignment. In addition, we provide a summary of tools and data available for gesture analysis, and describe speech-gesture interaction models and simulations in technical systems. This overview also serves as an introduction to a Special Issue covering a wide range of articles on these topics. We provide links to the Special Issue throughout this paper.
© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Messages can be encoded verbally or nonverbally. Although research on communication has traditionally focused on speech, recent years have witnessed a steadily growing interest in multimodality. Clear evidence comes from an increasing number of workshops attracting an international, interdisciplinary audience. For example, the GESPIN (Gesture and Speech in Interaction) conferences in Poznań (2009) and Bielefeld (2011), and the Gesture Workshop Series (GW) have focused on the technical modeling of manual gestures in human-machine interaction. The Audio-Visual Speech Processing Workshops (AVSP) have concentrated mainly on the technical aspects of multimodal facial communication, while the LREC Workshops on Multimodal Corpora, and the ISGS (International Society for Gesture Studies) conference series have each featured a broad spectrum of gesture research.

This strong interest is linked to the fact that clearly, accounting for verbal or textual information only, does not suffice to provide a full picture of human communication. Multimodality benefits speakers. For instance, when describing a cup we are searching for, we can use our hands to describe its shape and size while saying "It is about this big and is shaped like this". By using our hands, we avoid having to produce precise verbal descriptions of spatial dimensions. "Semiotic versatility" refers to the way that different modalities lend themselves to representing certain kinds of information better than others. Hands are better suited to expressing shape than speech, while the face best expresses emotions and attitudes. When communicating using their full multi-modal expressive potential, speakers can increase communicative efficiency, by simultaneously transporting complementary information, and foster robustness, by providing redundant information in various modalities.

The interplay between gesture and speech is highly adaptive to various situations. Speech may dominate when hands are needed for other tasks, while gestures probably take over in noisy situations. In any case, we often use the information in one modality to disambiguate, enhance or highlight the information in another modality. Kendon (2004) distinguishes two main functions of co-speech gestures, namely *substantial* and *pragmatic* gestures. The former contribute to the utterance content, while the latter help negotiating aspects of the situational embedding. This is done by conveying attitudes, levels of attention or agreement between the interacting parties, or by chunking the speech units into turns or information packages, thus guiding the discourse organization. Naturally, all of these aspects are of interest to basic and applied research.

Given the manifold functions and the complex interplay of the modalities, a full account of communication will need to describe and explain (a) the various types of functional, modality-specific information, and (b) how their interactions are constrained. Our hope is that this special issue will serve as an encouragement to an even deeper exploration of these questions, with a focus laid on the functional and temporal interactions and constraints existing between speech and gesture.

Given the need for more insight into the interplay between the two modalities, our goal is to promote discourse between gesture and speech research communities. A growing num-

ber of gesture researchers have broadened our understanding of the role of manual gesture in communication, but have so far rarely dealt with the more technical aspects of the gesture-speech interface. This is unfortunate, as technical systems allow us not only the development of working applications, but also provide a straightforward path to model simulation and evaluation. Likewise, more researchers on technical systems will probably profit from a better understanding of how gesture helps speech and language processing in humans (cf. Section 5.4). Also, gesture research may significantly benefit from an understanding of how prosody is linked to speech, as this link resembles many aspects of the gesture-speech relationship. Some researchers have even argued that intonation is the "gesture-like component of speech" (Tuite, 1993) or that it is part of a common production system co-expressive with the verbal stream (Bolinger, 1982; Kendon, 1972).

Besides giving an overview of our current understanding of the speech-gesture relationship, our main objective is to narrow the gap between speech and gesture research, and between perspectives on gesture taken in engineering vs. the humanities. Indeed, we feel that all these, largely overlapping, communities will profit from such a discourse, by making their models cognitively plausible, formally solid, transferable to real-world applications and empirically well-founded. In the remainder of this paper, we give an overview about how speech and gesture are linked temporally and functionally and discuss existing tools and methods for annotation, analysis and technical simulation. As the present article is also the introduction to a Special Issue covering a wide range of articles on this topic, we provide links to these throughout.

2. What are co-speech gestures?

According to Kendon (2004), a gesture is a visible action of any body part, when it is used as an utterance, or as part of an utterance.[1] We focus on those visible actions that are produced while speaking, namely, co-speech gestures. Their occurrence, simultaneous or concomitant to speech, has led to different views regarding their role in communication. Either, gesture is seen as an integrative, inseparable part of the language system (McNeill, 1992, 2005; Kendon, 2004), or speaking itself is regarded as a variably multimodal phenomenon (Cienki and Müller, 2008). Whatever the case might be, co-speech gestures vary in different respects. Originally McNeill (1992) differentiated them along, what he termed, *Kendon's continuum*. With a higher degree of conventionalization, gesture becomes less dependent on the co-occurring speech, with sign language being completely independent. Emblematic gestures, e.g. the "thumbs up" gesture, are conventionalized and language-specific, while co-speech *gesticulations* are less standardized

and work together with speech to accomplish communicative success. Later, McNeill (2005) further refined the idea and argued for a complex of several continua, namely

(a) Continuum 1: relationship to speech (obligatory presence of speech – ... – obligatory absence of speech)
(b) Continuum 2: relationship to linguistic properties (linguistic properties absent – ... – linguistic properties present)
(c) Continuum 3: relationship to conventions (not conventionalized – ... – fully conventionalized)
(d) Continuum 4: character of semiosis (global & synthetic – ... – segmented & analytic)

Gesticulations are placed on the left ends of these continua (co-speech, no linguistic properties themselves, not conventionalized, global meaning). In this special issue, the focus lies on gesticulations, as through them the full potential and limits of speech-gesture interaction can be examined. In the following sections, we refine our overview of manual gestures (cf. Section 2.1) and head gestures (cf. Section 2.2) respectively.

2.1. Gesturing with the hands

The gestural movements of the hands and arms are probably the most studied co-speech gestures. Based on the seminal work by Kendon (1972, 1980), they are usually separated into several *gestural phases*. A review is found in Bressem and Ladewig (2011):

(a) A *rest position*, a stable position from where the gesticulation is initialized,
(b) a *preparation phase*, during which a movement away from the resting position begins in order to prepare the next phase,
(c) a *gesture stroke*, which is typically regarded as obligatory and containing a peak of effort (directed at manifesting the communicative function) and a maximum of information density,
(d) *holds*, which are a motionless phases potentially occurring before or after the stroke, and
(e) a *retraction or recovery phase* during which the hands are retracted to a rest position.

Additionally, the point of maximal gestural excursion is often regarded as a gestural *apex* (see also Table 1 in Section 4.3). Several more detailed categories of gesture phases were proposed. These included the *recoil* phase (Kipp, 2004).

Gestures can be described in terms of their form, their semantic and pragmatic functions, their temporal relation with other modalities, and their relationship to discourse and dialogue context. Gut and Milde (2003) pointed out that a function-oriented gestural phase classification, such as the one by Kendon above, differs from form-oriented descriptions of gestural phases. In form-oriented

---

[1] This point of view excludes *self-adaptors*, usually understood as instances of touching self, scratching or neck massaging.