Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 209–232

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

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.

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-

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.09.008

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 conventialized, 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.

Table 1
Empirical studies reporting on the temporal relations between various gesture anchors and prosodic anchors.

Name/Author(s)	Gestural anchor term	Gestural anchor definition	Prosodic anchor	Temporal relation
Loehr (2004, 2012)	Gesture apex (of effort)	"peak of a stroke", defined differently for pointing vs. other gestures	Pitch accent	Mean pitch accent lead of 17 ms, Std. Dev = 341 ms
Leonard and Cummins (2009, 2010)	Beat gesture apex, peak velocity	Point of maximum extension, peak velocity of the extension phase	Vowel onset, pitch peak on a stressed syllable, the P-center	Apex aligned with peak pitch with relatively lowest variability, velocity peak aligned with either P-center or vowel onset
Roustan and Dohen (2010)	Pointing, beat gesture apex	extension of the index finger, end of a downbeat	Articulatory targets, *F*0 and intensity peaks	Pointing apex aligned with articulatory targets, downbeat with prosodic peaks
Rochet-Capellan et al. (2008)	Pointing gesture apex	Onset of the finger target alignment period	Jaw opening gesture in a stressed syllable	Apex synchronized with the gesture in 'CVCV words
Yassinik et al. (2004), Shattuck-Hufnagel et al. (2007)	Gesture hit	The point of an abrupt stop of movement, often followed by a change in direction	Pitch accented syllables, stressed syllables	90% of gesture hits occurring on a pitch accent
Jannedy and Mendoza-Denton (2005)	Gesture apex	"peak of a stroke" as in (Loehr, 2004) or "gesture target"	Pitch accent	95.7% of apexes co-occured with a pitch accent
McClave (1991, 1994)	Beat gesture phases	Downbeat and upbeat	Nuclei of tone groups, stressed syllables in multisyllabic words	Downbeats co-occur with nuclei, both down- and upbeats with stressed syllables in multisyllabic words

approaches, it is common to break down a gesture into several morphological features, e.g. the handshape, location, hand direction and movement type (cf. Section 5.1). The frame of reference for form feature descriptions is usually the *gesture space* (McNeill, 1992), which organizes the space in front of the speaker's body into positions, regions and directions. For example, gestures in *central gestural space* can be distinguished from those in *peripheral gestural space*. The majority of gestures are performed in central gestural space, a sphere-shaped area in front of the speaker's upper part of the body, below the neck and between the shoulders and elbows. Gestures occurring outside this area are said to be produced in peripheral gestural space and are believed to capture the addressee's visual attention.

In addition to primitive form features, it is useful to distinguish different representation techniques (Kendon, 2004) or gesture practices (Streeck, 2008) like *shaping*, *drawing*, *modeling*, or *acting*. These techniques represent more complex patterns of gestural performance. They constrain the more primitive features, which helps to interpret correlations between them, and to analyze their meaning-form mapping (Bergmann and Kopp, 2010). Studies by Priesters and Mittelberg (2013) support these results and suggest that the usage of gesture space and form, as a function of meaning and context-related factors, reveals speaker idiosyncrasies, probably reflecting speaker-specific individual communicative styles, cognitive and verbal skills, or personality traits (Hostetter and Alibali, 2007; Hostetter and Potthoff, 2012). *Aktionsarten*, as a means to classify the internal structure of events into types such as states, activities, accomplishments and achievements (Vendler, 1967) were used to analyze the relationships between verbs and accompanying gesture strokes by Becker et al. (2011). The results revealed a systematic timing variability of the verbs and gestural strokes when interpreted as co-expressive Aktionsarten.

Generally, as manual gestures serve various functions in communication, it is often useful to characterize and classify them with respect to their semantic function (alone, or along with formal features), e.g. using the classification by McNeill (1992):

(a) *emblematic gestures* bear a conventionalized meaning ("thumbs up");
(b) *iconic gestures* resemble a certain physical aspect of the conveyed information, e.g. they may convey the shape of a described object or the direction of a movement;
(c) *metaphoric gestures* are iconic gestures that resemble abstract content rather than concrete entities (McNeill, 1992; Cienki and Müller, 2008);
(d) *deictic gestures* point out locations in space, with space often being of a conceptual rather than concrete nature;
(e) *beat gestures* are simple and fast movements of the hands (also called *batons* (Ekman and Friesen, 1972)). Rather than directly conveying meaning, they refer to the process of speaking itself by synchronizing with prosodic events in speech. They have been found to contribute to the perceived prominence of temporally aligned speech, and can function in the sense of *Audiovisual Prosody* (cf. Sections 2.2 and 4.2).

It is important to note that this classification should not be understood as defining distinct categories. McNeill (2005) convincingly argued that a simple functional classification of gestures is usually misleading. Due to the multi-

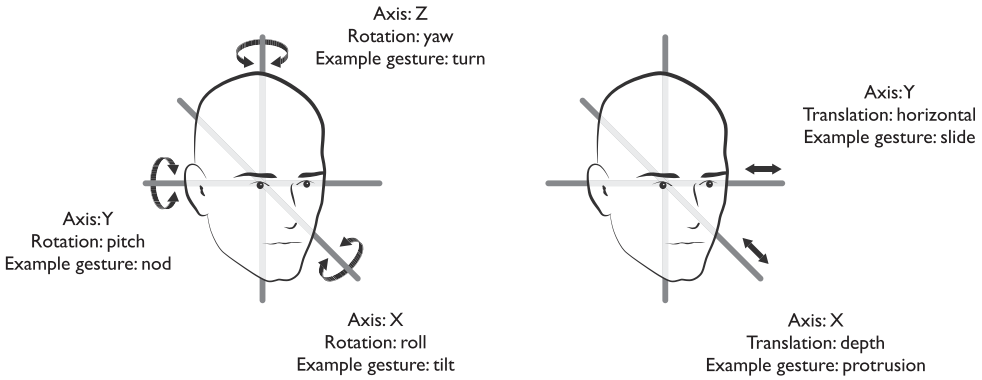

Fig. 1. Schematic overview of rotations and translations along three axes, as well as example movements most frequently used in communicative head gesturing.

faceted nature of most gestures, he prefered a dimensional, rather than category-based characterization of gestures, with dimensions including *iconicity*, *metaphoricity*, *deixis*, *temporal highlighting* (beats), and *social interactivity*. This acknowledges the fact that the majority of gestures can be characterized along several of these dimensions, e.g. when a pointing gesture also depicts the direction of a movement, or when a beat is superimposed onto the stroke onset of an emblematic gesture (Tuite, 1993).

2.2. Gesturing with the head

Due to their dynamic variability and multidimensionality, it is challenging to identify characteristic kinematic parameters contributing to classifiable head gesture "segments" or patterns (Altorfer et al., 2000; Kousidis et al., 2013). However, some generalizations about basic head gesture forms and related functions can be made.

Communicative head gestures utilize movements around three main axes of rotation (yaw, roll and pitch rotations around the Z, X and Y axes respectively) as well as typically two linear displacements (translations) along the Y (inter-aural) and the X axes (naso-occipital). The most common rotations and translations used in co-speech head gesturing are presented schematically in Fig. 1. Mathematical conventions for 3D spatial coordinates are used in Fig. 1, as done in biomechanical and physiological studies (Yoganandan et al., 2009; Kunin et al., 2007).[2] The figure presents simple movement forms. However, it is important to remember that these patterns vary greatly in their exact kinematic realizations (angles, extent), as well as overlap with other movements. The pitch rotation in the up-down direction (lowering and raising about the Y axis or flexion/extension) is especially often used in conversation. The movements associated with this rotation are usually called head nods and head jerks or upstrokes. Recent analyses of multimodal corpora report that head nods predominate in gesturing when a listener role is assigned to the participants (81.5% of head movements were nods in Włodarczak et al., 2012 in a German active listening corpus) or in free conversation (56% in a multimodal corpus of Japanese in Ishi et al., this issue). Lateral movements around the Z axis are usually labeled as turns and shakes, while angular displacements about the X axis (lateral bending) are called tilts. Linear displacements (translations) are less frequently used, however, protruding and retracting the head along the X axis may correlate with displays of surprise and attention. We also proposed the inclusion of displacement along the horizontal axis, based on observations of German corpus data, as reported in Kousidis et al. (2013).

Apart from displacements, several other kinematic parameters are important for identifying head movement patterns. The motion type examples discussed above suggest that the *direction* along the same axis distinguishes between communicative forward and backward head movements, i.e. between a head nod and a head jerk (downstroke vs. upstroke). *Cyclicity* plays a role in distinguishing types of lateral movements such as (repeated) shakes, where the head acts as a pendulum (Heylen et al., 2008), as opposed to single turns of the head to one side or the other. Hadar et al. (1985) differentiated between linear and cyclic kinematic forms, equivalent to e.g. single and multiple nodding bouts and associated them with turn taking signals and responses to questions, respectively. Single nods vs. multiple nods were suggested to vary with feedback function in a language dependent manner (Hadar et al. (1985) for English, Cerrato (2007) for Swedish, Włodarczak et al. (2012) for German, Ishi et al. (this issue) for a detailed account of head movement-dialogue act relationships in Japanese).

Continuous variation along parameters associated with the *intensity* of head gesturing, such as movement frequency (movement rate) and amplitude (degree of displacement) may contribute to the nuancing of several main

[2] Note that in computer graphics it is customary to represent the vertical axis as Y, the horizontal as X and the depth axis as Z (Buss, 2003). It is also a convention most likely to be used in motion tracking system manuals as well.

functions (cf. *motion qualifiers* modifying kinesic structures, i.e. intensity, range and velocity, proposed by Birdwhistell (1970)). The nuancing is associated primarily with giving feedback, turn-taking and audiovisual prosody. Hadar et al. (1985) noted that floor grabbing cues were usually expressed with wide and linear head movements (e.g. high amplitude, single nods), while synchronization with pitch accented syllables in the interlocutors speech occurred in cases of narrow, linear head gestures, e.g. low amplitude single nods. Rosenfeld and Hancks (1980) suggested that smaller, single nods often function as typical backchannels. While, according to Bousmalis et al. (2012), large amplitude, repeated nods are characteristic of affirmative meanings. The lateral dimension of head movement seems to exhibit several possible coherent combinations of kinematic features characterizing different patterns. Heylen et al. (2008) proposed a differentiation between cyclical shakes, head sweeps, slow head moves, head repositionings and other lateral movements, where all classes could be related to different functions.

Attitudinal and emotional information is easily communicated by head movement intensity as well. As Hadar et al. (1985) noted, a single, rapid and sharp head nod may signal impatience, a moderately fast, repeated nodding could be a simple sign of agreement. Qualitative parameters associated by some authors with contextual meanings or attitudinal information, such as *jerkiness* (Poggi et al., 2010) or *fluidity* (Hartmann et al., 2006) of head movement need to be better described in terms of, possibly, non-linear kinematic parameters for a more optimal physical characterisation in the future.

Hadar et al. (1983) attempted a spectral classification of speech-related head movement, independent of the usual type labels listed above. A well-known result of this report is the incessant head movement found during speaking turns in his data: 89.9% of the time. Distinct spectral features were proposed to be characteristic of communicative head gestures as opposed to mere physiological tremor (Hadar et al., 1983) and movement caused by body posture shifts. The frequency ranges for communicative head movement were found to lie between 0.2–7 Hz.[3] Within this range, movement amplitude was inversely correlated with frequency, i.e. the faster the movement, the more limited its excursion in space, with the exception of postural shifts, occurring typically with a high amplitude and high frequency movement.

Apart from the repetition of the same movement, different head movements can be concatenated and layered in uninterrupted sequences (Heylen et al., 2008). The rotations and translations discussed above are simple examples of classifiable patterns, describing typical *exclusive movement changes* (Altorfer et al., 2000), i.e. movements that often appear on their own and are associated with relatively specific functions. On the basis of a corpus of spontaneous dialogue reported on in Kousidis et al. (2013), it is apparent that approx. 30% of uninterrupted, coherent head movement "phrases" (called Head Gesture Units) contain up to 10 different exclusive movements in concatenated sequences, compared to single or repeated instances of the *tilt*, *turn* or *nod* type. Additionally, Kousidis et al. (2013) evaluated a manual head movement annotation scheme based on kinematic parameters, as described above. They found that, for instance, annotators very often confused tilts, shakes and nods with turns, possibly because the former were overlaid on a broader head turning movement (cf. Heylen et al. (2008)). Head tilts and nods also blended frequently.

After having concentrated on the characteristics of gestural form, the following section will deal with the functions of co-speech gestures in facilitating communication, meaning construction and conceptualization, as well as language planning and production.

3. How do speech and gesture interact?

Various interactions have been found to exist between gesture and speech. These arise generally when cooperating in organizing communication (cf. Section 3.1) and in more specific ways, e.g. when adding representational content to a verbal message redundantly or non-redundantly (cf. Section 3.2), or when spelling out concepts multimodally in speech production, thus revealing language production mechanisms (cf. Section 3.3).

3.1. Gestures in the communicative context

Listeners, as well as communicative interactions as a whole, can benefit from gestural information. Kendon (2004) suggested that the "pragmatic" function of gesture can be *modal* when expressing a speaker's stance, *performative*, when referring to the ongoing speech act or the interpersonal move made, *parsing*, when highlighting stretches of speech in a demarcative function, or *interpersonal*, when referring to the speaker's role or organizing the sequencing of turns.

Speech and gestures have been described as following a pragmatic synchrony rule, meaning that, when produced together, the modalities always fulfill the same pragmatic function (McNeill, 1992). As already noted above, however, the modalities also fulfill complementary pragmatic functions, as in when a beat gesture or a head nod highlight information that is delivered verbally. Likewise, we find especially close semantic and pragmatic coordination between head movement and the spoken signal, as well as between head movement and gaze. These relationships exploit the economy and degrees of expressiveness enabled by the simultaneous activation of channels in the same

[3] The range of frequencies proposed to define communicative head gestures as reported in Hadar et al. (1983), is located comfortably within the range of head rotation frequencies compensated for by the vestibuloocular reflex (VOR). VOR provides eye rotation of equal and opposite magnitude to head rotation, stabilizing gaze in space while the head is moving either voluntarily or in locomotion (Tomlison et al., 2000).

bodily area. This physical relationship may explain the evident and uncontroversial close meshing of several functions in head gestures: the interactive, the pragmatic and the semantic, as well as the attitudinal and emotive. Manual gestures in comparison, *seem* to act relatively more independently and to have a differentiated capacity to represent propositional content, adding a mode of expression that is often faster and more effective than speech. Findings by Morrel-Samuels and Krauss (1992) precipitated the research focus that related manual gestures to cognitive models of speech production and lexical access processes. Recently, however, the interactive dimension of manual gesture use has begun to gain ground.

The fact that a principle function of gesture is to aid communicative needs is underscored by findings showing a decrease in gesturing when there is no visibility between the speaker and the listener. Also, gesturing appears to change as a function of the type of communicative situation (cf. Bavelas et al. (2008) for an overview). Another indicator for the communicative function of gesture is that visibility has an effect on the types of gestures produced. When there is visibility, gestures are larger (Bavelas et al., 2008) and often have clear interactive functions such as the regulation of turn taking (Bavelas, 1994). That said, non-referential beat gestures (Alibali et al., 2001) and non-obligatory iconic gestures (de Ruiter et al., 2012) were found under both visibility and non-visibility conditions. This implies that some gestures have a function predominantly facilitating speech production (cf. Section 3.3), while others' main function is of a more communicative nature.

de Ruiter et al. (2012) suggest that speakers often mix speech and gesture redundantly in the initial stages of dialogue in order to decrease the risk of misunderstandings. The explanation for such an effect would be the desire to minimize joint effort in communication (Clark and Wilkes-Gibbs, 1986). In other words, speakers opportunistically choose the most effective solution to repair communication problems. Therefore, if gesture can offer a quick way to disambiguate or add meaning, it is likely a speaker will use it.

Holler and Beattie (2003) proposed that gesture often arises as a response to the immediate communicative demands that a situation poses, at least in case of a repair. In the case of lexical ambiguity, the quickest solution to resolve the communication problem is chosen, i.e. typically a disambiguating gesture is performed (Clark and Wilkes-Gibbs, 1986). Listener's needs are clearly recognised and acted upon. Holler and Stevens (2007, 2009) also showed how the depth of *common ground*, i.e. the degree of an interlocutor's belief about mutual understanding (Clark and Brennan, 1991), can affect gesture and speech, and provided further evidence that utterances are multimodally "designed" for the benefit of the recipient, depending on the level of shared knowledge (de Ruiter et al., 2010).

Establishing common ground is also one of the most prominent pragmatic function of head gestures. Interlocutors express ongoing attention and understanding, and display appropriate listening behavior in dialogue by giving feedback (Bevacqua, 2009; Heylen et al., 2011) in the form of nods, and sometimes, but not always, co-occurring with verbal feedback signals such as *yes* or *u-huh* (Allwood and Cerrato, 2003). Dialogue partners also elicit feedback with head gestures (McClave, 2000; Goodwin, 1981). Feedback produced by means of head gestures (and smiling) is hypothesized to be a particularly good backchannel, since it provides simultaneous information about the success of communication, without disrupting the speech of the interlocutor (Włodarczak et al., 2012; Heldner et al., 2012).

Head movement can also signal turn claims (Duncan, 1972; Hadar et al., 1984), turn continuation and changes from direct to indirect discourse through postural head shifts (McClave, 2000), as well as topic and narrative changes (Kendon, 1972).

In addition, the head can be used for deixis by protruding directly towards an object. This motion can signal discourse related abstract deixis, as in reference to an alternative topic. It can point to elements in a list (Kendon, 1972; McClave, 2000; Heylen, 2005, 2008). It can also indicate the referential use of space (McClave, 2000), when marking contrast between topics or objects by deixis in space (Heylen, 2006).

Given these points, it is important to remember that head gestures often need to be interpreted in their co-occurring linguistic context, such as preceding or overlapping feedback expressions, and/or simultaneous multimodal context (co-occurring facial displays, gaze behavior or hand gestures (Poggi et al., 2010; Rosenfeld and Hancks, 1980)). Heylen et al. (2008) report on several studies where head movements co-occurring with facial expression resulted in a different functional interpretation. The head tilt is thought to express disbelief or lack of understanding, although when adding e.g. a facial expression, it can express interest and/or surprise. Similarly, a nod with a frown can express "dislike". Bevacqua (2009) found evidence that head nods generated by a listening agent were interpreted as "agree" and "understand" by human study participants; however, when combined with a smile, they were interpreted as "like" and "accept". Beskow et al. (2007) investigated the weight of particular cues in agents, for example "laughter as agreement" and found similar correlations.

As mentioned earlier, many principal functions are shared by two closely located modalities. The organization of turn taking (Jokinen et al., 2010) as implemented by head gesturing is often discussed in the relation with eye gaze. Consider also attention displays. Mutual gaze often functions as a prerequisite for communication, opening or closing the *communication gate* for other modalities, signaling basic levels of attention. Bavelas et al. (2006) reported on a higher number of responses from a listener within so-called *gaze windows*. The presence of visual feedback, primarily head gestures, is closely linked to mutual gaze and correlates with active and attentive listening. Truong et al. (2011) found that in face-to-face communication, the effect of pitch contours in cueing backchannels was less

significant than mutual gaze. Specifically, head gestures were found to be significantly timed with mutual gaze.

3.2. Transporting meaning

The parallel use of gesture and speech gives the speaker and the listener access to *complementary* or *supplementary* semantic qualities that are not present in the accompanying speech (Goldin-Meadow et al., 1993). Supplementary gestures denote information not referred to simultaneously in speech, e.g. when a child points to a candy while saying "I'm hungry". Complementary gestures denote additional aspects of information referred to in speech, e.g when saying "The ball is this big" and simultaneously indicating its large size. Goldin-Meadow (1999) and Goldin-Meadow et al. (2001) showed that the usage of these additional pieces of information enhances communication and lowers cognitive load in both comprehension and production. Rowbotham et al. (this issue) showed that gestures were especially suited to conveying supplementary information when referring to qualia such pain sensations. Further support for the idea of gesture facilitating speech planning and production comes from Alibali et al. (2000), who found a positive effect of gesture on children's speech planning. Iverson and Goldin-Meadow (1998) reported on blind speakers gesturing when addressing other blind listeners. They thereby provided evidence that gesturing is not purely listener oriented but an integral, probably facilitating, aspect of speech production (also cf. Section 3.3).

Specifically, gestures may support the expression of abstract concepts. Metaphoric gestures, for example, help speakers convey complex representations by frequently providing complementary information about mental states, mathematical concepts, metaphysical phenomena, spatial imagery in discourse structure, and conceptualizations of language and communication. These are known as *conduit gestures* in the cognitive linguistics literature (McNeill, 1985; Cienki and Müller, 2008). Once the meta-level of abstract thinking is sufficiently developed, people very often make use of gestures to depict the image of the abstract concept they have in mind (Alibali and DiRusso, 1999; Alibali et al., 2000). Perhaps therefore, the analysis of metaphoric gesture use has proven its fruitfulness in, for example, teaching (Roth, 2001; Goldin-Meadow, 2003).

In some contexts, even beat gestures can convey complementary abstract meaning relating to dialogue structure. In such cases, they may refer to words with which they are synchronized, or to a different point in the conversation, thus appearing to help structure discourse (McNeill, 1992). Pointing gestures very often accompany topic shifts by orienting the listener towards contrasting spaces, where one space refers to a previous topic and the other to the next (McNeill, 1992). In this case, complementary, metaphorical references mix with the main interactive function of pointing and beat gesture use, as discussed in the previous section.

In addition to being complementary or supplementary, gestures can be *redundant* relative to the accompanying speech, as when saying "The ball looks like an orange" while also gesturing size and shape. Often, the main distinction can be reduced to redundant vs. non-redundant gestures, the latter comprising complementary and supplementary gestures. However, the meaning of a gesture and the meaning of speech are *always* of a fundamentally different nature due to their different semiosis. As such, it is questionable whether gestures can truly be thought of as being strictly redundant (McNeill, 2005). The de-facto contemporary standard approach to examining this complex relationship lies in a deep semantic analysis. A gesture is assigned a single propositional meaning or semantic features, which are then individually compared to the semantic features present in the speech signal. Analyses adopting this approach have shown a roughly even proportion between redundant and non-redundant gestures (Bergmann and Kopp, 2006).

Similarly to manual gestures, most form-function pairing of head gestures are non-conventionalized. Other than manual gestures, head gestures tend to be regarded as predominantly serving pragmatic rather than semantic functions (McClave, 2000). Manual gestures are clearly more suited to propositional content expression than head movements due to the natural biomechanical constraints of the respective body parts. As such, the most common semantic function for head gestures involves positive and negative responses associated with head nodding and shaking,[4] as well as gestures such as lateral sweeping with the head to mean "everyone" or "anything" (Goodwin, 1981; McClave, 2000). McClave (2000) also observed the use of lateral shaking intensification to mean "great!" "really..?", "exactly", "totally" as well as to signal uncertainty. These meanings, however, very often coincide with facial and/or prosodic expressions that participate in disambiguating the meaning within a complex multimodal signal, especially in the absence of verbal disambiguation.

Poggi et al. (2010) suggests that despite the inherent, context-dependent, polysemous nature of head movements, it is still possible to identify *semantic cores* common to many particular uses. These cores are thought to depend on speakership roles. For head nods, in speaking turns, the semantic core of "importance" (prominence) is apparent, while what is inherently communicated in listening turns is "acceptance". Similarly, Kendon (2003) suggests that head shakes, in all contexts, share a core meaning of negativity.

However, the possible iconic or universal nature of the relation between e.g. nodding and agreement cannot be postulated, since some well-known reversals exist. In Bulgarian (Jakobson, 1972), where a head upstroke means "no" and in some cultures of the Mediterranean (Greece,

[4] Compare this to extending the hands towards the interlocutor with palms towards her and waving the hands sidewards, or finger wagging as a symbolic expression of disagreement and refusal (Jakobson, 1972).

Southern Italy, the Balkans) where throwing the head back is associated with negation (Abercrombie, 1954; Jakobson, 1972). The *head bobble* or *waggle*, referring to repeated, side-alternating head tilting, is a characteristic head gesture in India, unknown in Western cultures, used to express backchanneling, friendliness and acceptance. Clearly, even basic meanings and communicative functions are associated with gesture forms embedded in a specific cultural context.

3.3. Producing gesture and speech

Numerous models have been formulated to provide possible explanations for the coordinations observed in human speech and gesture. The distinction between redundant and non-redundant gestures has led to different points of view regarding the production and perceived facilitating effect of speech-accompanying gesture. One point of view on speech-gesture interaction proposes that gestures tend to reflect what is simultaneously produced verbally in what has been called the *hand-in-hand hypothesis* (So et al., 2009). An alternative view is that of a *trade-off relation* between gesture and speech production. In this view, speech and gesture complement each other in transporting information, with the less costly production channel as more dominant (cf. de Ruiter et al. (2012) for an overview).

More detailed accounts describe interactions between gesture and speech at different stages of the production process, e.g. conceptualization or lexicalization. Numerous researchers have suggested that gestures, especially representational gestures (Krauss and Hadar, 1999), play a direct role in speech production through priming the lexical retrieval of words. This view has been termed the *Lexical Retrieval hypothesis*. The hypothesis was based on research arguing that (1) gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewelyn, 1969; Butterworth and Beattie, 1978), and (2) that the inability to gesture can cause verbal disfluencies (Dobrogaev, 1929). However, other studies have reported conflicting findings (Nobe, 2000), e.g., more gesturing in the absence of filled pauses (Christenfeld et al., 1991) or no significant effect of gesture inhibition on verbal delivery (Hoetjes et al., this issue).

The implications of the Lexical Retrieval hypothesis (LR, henceforth) are often contrasted with the *Information Packaging hypothesis* (IP, henceforth) by Kita (2000), Alibali et al. (2000), Hostetter et al. (2007). This hypothesis is based on findings showing that speech and gesture interact at an earlier stage when information is packaged, organized, and distributed across the modalities. That is, this view assumes that gesture is involved in the conceptualization of information for speaking. Empirical support comes from studies showing increased gesturing when the conceptualization load for speaking is high (Kita and Davies, 2009), when speakers have strong visuo-spatial skills and weak verbal skills (Hostetter and Alibali, 2007), or when they introduce new information into the dialogue (Bergmann and Kopp, 2006). Likewise, speakers of different languages package information differently for their target language in a thinking-for-speaking process (Slobin, 1996) and are found to follow the language-specific packaging in their gestures (Kita and Özyürek, 2003) as opposed to making overly non-redundant gestures.

It can be said that the two hypotheses differ in the "depth" of the connection between speech and gesture where the LR associates it with surface spoken forms, and IP with a deeper level of organizing and encoding spatio-motoric information (Alibali et al., 2000). Analogous to LR, where problems in lexical retrieval are associated with increased gesture activity, in the case of IP, difficulty in conceptualizing information predicts an increased use of gesture. It is important to note that causes and effects in gesture and speech production are still not well understood. For example Alibali et al. (2000) indicated two versions of the LR hypothesis. The first argued that gestures derived from spatially encoded knowledge, and facilitated access to lexical entries (Krauss and Hadar, 1999). The second that gestures derived from lexical entries and facilitated retrieval of phonological forms (Butterworth and Hadar, 1989).

A number of different models have been proposed to explicate the different assumptions about how speech and gesture interact during production. The LR hypothesis has led to a model (Krauss and Hadar, 1999) postulating that speech production and gesture production are rooted in separate memory representations (visuo-spatial and propositional, respectively). Semantic features are thought to be activated and processed by the respective system without any coordination. Cross-modal interaction is believed to result when features happen to be processed by both systems, and at a later stage when a selected gesture can prime words during lexical retrieval.

Other models allow for more interaction, in following with McNeill's *Growth Point theory* (McNeill, 1992, 2005) which states that speech and gesture are in fact inseparable parts of one and the same self-organizing process. In the *Sketch Model* (de Ruiter, 2000a), interactions are assumed to be located in a shared conceptualizer responsible for any multimodal coordination (semantic, pragmatic, temporal). The Sketch Model provides explicit accounts for different kinds of gestures. Kita and Özyürek (2003) extended this idea by postulating two different subcomponents, a message generator for conceptualizing speech and an action generator for producing manual actions and communicative hand gestures. In this view, gestures are based on spatio-motoric thinking, but are also thought to interact with speech through bi-directional links between the two subcomponents. This interaction is assumed to include translation and matching of modality-specific content organization via an interface representation (Kita, 2000) and to converge towards an information packaging and distribution that is readily verbalizable.

Hostetter and Alibali (2008) proposed that gestures were simulated actions in the mind of the speaker (the *Gesture as Simulated Action* model). Based on an embodied cognitive

view, language and imagery are assumed to evoke mental simulations that spawn motor activations. Depending on a speaker- and situation-specific adaptive threshold, these activations may be executed leading to overt gestures. This model focuses on action-related (pantomimic) gestures that may directly reflect internal simulations during the moment of speaking. Finally, in a recent cognitive account, Kopp et al. (2013) and Bergmann et al. (2013) have proposed a model that grounds conceptualization in the formation of multimodal messages in working memory. This view assumes that visuo-spatial and propositional representations are activated and linked dynamically during thinking for speaking, and that the simultaneous operation of the speech and gesture production system on the memory gives rise to a memory organization that is (1) temporally stable, (2) linguistically-shaped, and (3) verbalizable. Conceived as an example of spreading activation, this model succeeds in accounting for many of the findings on the packaging and distribution of information across speech and gesture with different cognitive or linguistic constraints. As far as the multifaceted nature of gesture functions and their interaction with speech in communication is concerned, contemporary production models appear to remain restricted in scope. For this reason, Ferré (this issue) suggests an extension of existing gesture production models using an additional pragmatic component (cf. also Section 3.1).

4. The importance of temporal coordination

Given the intimate connection of gesture and speech in communication, their *precise* temporal coordination stills bears many open questions and is the subject of ongoing debate. Regardless, this factor needs to be taken into account by unified speech and gesture production models. It also requires treatment in technical systems such as virtual or embodied artificial agents. However, when trying to relate the empirical results of studies on speech-gesture synchrony to models of gesture production, it seems that the ideas inherent to the McNeillan tradition of a common source for gesture and speech have not been directly translated into predictions involving the coordination of speech and gesture in time. If gestures are planned from the same source, as suggested by the Sketch Model's conceptualizer (de Ruiter, 2000a), synchronization should be planned beforehand and later imposed on the common output in one go. Both the Interface model (Kita and Özyürek, 2003) as well as Hostetter et al. (2007) suggest that gesture comes from a source separate from the speech production process, namely the *action generator*, predicting that gestures exhibit properties of practical action. Gesturing consequently exploits the affordances of referent objects and is influenced by them. Still, there is no explicit prediction about coordination in time for the Interface Model. However, it is possible that by regarding gesture and speech as planned from within two different sources, we gain degrees of freedom enabling a dynamical adaptation of speech and gesture timing, helping to reach communicative goals flexibly and promptly address disfluencies (Rusiewicz et al., this issue).

In the remainder, of this section, we give an overview of current theoretical and model oriented views about the temporal interaction of gesture and speech without referring to signal-based analyses of this relationship (cf. Section 4.1), in Section 4.2, the special role of prosody in the temporal interaction between gesture and speech is discussed. In Section 4.3, the various attempts at measuring the speech-gesture interaction, usually building on the close prosody-gesture relationship, are introduced and critically discussed.

4.1. Temporal interaction of gesture and speech

At first glance, gesture and speech may be coupled less directly than, e.g., prosody and speech, as both originate in very different physiological systems. However, some views and findings suggest a close connection between both, especially in production. The assumption of a physiological link has been based on the view that "the mutual co-occurence of speech and gesture reflects a deep association between the two modes that transcends the intentions of the speaker to communicate" (Iverson and Thelen, 1999). The conceptual link in production was already discussed in Section 3.3. This point of view assumes a strong developmental sensorimotor link between the movements of the hands and mouth shaping an interdependence of both that stays throughout a person's lifespan. Iverson and Thelen (1999) argue that this interdependence can be seen in both neurophysiological data (also see Loevenbruck et al. (2009)) and in the motor coordination of hands and mouth (also see Gentilucci and Volta (2007)).

In a debate about the degree of linkage or independence, McNeill (1985, 1987, 1989, 1992) argued that the synchrony between gesture and speech provided evidence for the cognitive interdependence of the two modes. To support this hypothesis, he pointed to structural and functional parallels and provided evidence from acquisition showing the synchronized development of speech and gesture.[5] Another argument is founded in how different types of aphasia may manifest themselves similarly in gesture and speech. McNeill (1992) suggested three rules of synchronization between gesture and speech.

(a) The phonological synchrony rule (henceforth PSR) predicts that a gesture stroke should occur before the most prominent syllable.
(b) The semantic synchrony rule predicts that co-occuring gestures and speech relate to the same *idea unit*.
(c) The pragmatic synchrony rule predicts that co-occuring gestures and speech have the same pragmatic function.

[5] An overview on the topic of gesture acquisition is provided by Esteve-Gibert and Prieto (this issue).

Systematic cases of PSR were detected in studies on co-speech pointing gestures, where stressed syllables of demonstratives synchronized with deictic pointing gestures (e.g. Esteve-Gibert and Prieto, this issue; Levelt et al., 1985; Rochet-Capellan et al., 2008). Feyereisen (1987), Butterworth and Hadar (1989) challenged the straightforward view of speech-gesture synchrony. They argued that the precise nature of the temporal relationship between gesture and speech was still far from clear, as were the motivations for their temporal relationship. Generally, it is accepted that the onset of a gesture phrase precedes the onset of speech (Kendon, 1980; Schegloff, 1984; Morrel-Samuels and Krauss, 1992; Nobe, 2000; de Ruiter, 2000a). It is possible that the speech-gesture synchrony that McNeill evoked as one of the fundamental sources for the concept of the growth point, could be interpreted in less strict terms, as proximity rather than simultaneity.

Nonetheless, most often, the temporal lead of gesture relative to speech comes to be understood as evidence of synchrony. The same term is used when considering temporal correlations between points and phases of gesture related to the *lexical affiliate* (Schegloff, 1984). Similarly, in studies concerned with an even tighter relationship between gesture and speech landmarks, correlated with some notion of effort in gesture movement and prosodic structure, the term synchrony is used as well. We consider precedence relations inherent to the former two understandings of synchrony, and leave the more precise temporal anchoring and coordination problems to Sections 4.2 and 4.3.

Utterances and intonational phrases as well as prominent syllables and lexical affiliates are linguistic constituents that in fact correlate temporally. Additionally, given that gesture types can be more propositional, i.e. semantic (iconics) and more kinetic, i.e. prosodic (beats, deictics), it is difficult to differentiate which linguistic levels and constituents (words? or prominent syllables of words?) relate more systematically to which gesture semiotic type when discussing temporal relations. The interactions are many. Continuing with precedence, McClave (1991) showed that in case of complex gestures where several movements appear in succession, gestures were compressed and fronted to all finish before a stressed syllable. For head gestures, Dittmann and Llewellyn (1968) and Włodarczak et al. (2012) found that head movements used as feedback bimodally with an affiliated spoken feedback expression typically preceded the affiliate by about 200 ms for different sets of data in English and German. These precedence relationships were supported in several studies for different languages, e.g. by Ferré (2010) for French, Karpiński et al. (2009) for Polish and Chui (2005) for Chinese.

The above phenomena have been associated with gesture and speech production models (Fig. 2) in different ways, both in an attempt to substantiate the models and to explain the phenomenon itself. For example, McNeill's growth point theory (McNeill, 1987) assumes that thinking is imagistic, and speech and gesture both arise from the same imagistic source and the same computational stage. According to McNeill, the "gesture lead" phenomenon arises, since gesture, unlike speech, does not require linguistic processing. Morrel-Samuels and Krauss (1992) proposed that the common origin of gesture and speech is located on the pre-semantic level of communicative intention which activates both abstract propositional representations and motoric representations. Gesture anticipates speech because search times while accessing motoric representations are shorter than for semantic representations. Moreover, motoric representations are more differentiated than the restricted size of any lexicon. Morrel-Samuels and Krauss (1992) in fact found that the interval duration by which gesture precedes speech, as well as the duration of the gesture, appeared to be a function of how familiar the lexical affiliate was to the speaker.

Loehr (2012) also found that intermediate intonation phrases aligned with gesture phrases. In phonology, it is usually assumed that a full intonation phrase may consist of one or several intermediate phrases which are delimited prosodically by boundary tones (Beckman and Pierrehumbert, 1986). Loehr produced evidence that speech and gesture showed a certain pragmatic, temporal and structural synchrony. He suggested that the intermediate phrase constituted the minimal size of a cognitive package in which both intonational and gestural expression cohered.

In their study on beat gesture perception, Leonard and Cummins (2010) found some evidence for McNeill's PSR. Listeners systematically detected mismatches in co-speech beat gestures if they occurred only 200ms later than naturally produced ones. Gestures synthesized earlier than their

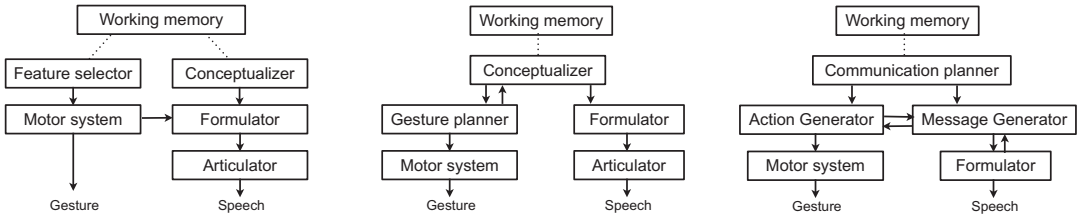

Fig. 2. Schematic overview of three different models that can account for interaction between speech and gesture production From left to right: Krauss and Hadar (1999), de Ruiter (2000b) and Kita and Özyürek (2003).

natural counterparts were often perceived as perfectly natural. Using ERP techniques, Özyürek et al. (2007) and Habets et al. (2011) found some support for the rule of semantic synchrony. They argued that in order to perform semantic integration, speech and gesture information should not be produced with noticeable asynchronies. They also postulated that a clarification as to which parts within the gestural and speech stream were in fact synchronized, and how, is needed in order to examine the temporal link between gesture and speech in more depth.

Schegloff (1984) suggested that typical affiliates for gestures can be a series of stressed syllables, typically co-occurring with a series of beat gestures. However, he claimed that this "on stress" production was a secondary ordering mechanism used to establish meaningful links between gesture and lexical or pragmatic components in dialogue. He argued that iconic gestures were typically produced *before* their lexical affiliate.

Other studies have looked at synchrony from the affiliation point of view as well, looking for lexical or semantic rather than prosodic hooks for co-speech gesture synchronization (Bergmann et al., 2011; Chui, 2005; Ferré, 2010; Harrison, 2013; Nobe, 2000). Treffner et al. (2008) studied the influence of gestural timing on the perceived prominence of a lexical affiliate. Their results suggest that a gesture needs to be realized slightly before the affiliate in order to strengthen its perceived prominence. Urban (2011) found that the point of highest prosodic prominence within a lexical affiliate served as anchor point for the gestural apex. Kim et al. (this issue) found differences in synchronization between head and eyebrow gestures and prosody, partly determined by information structure. Similarly, Ishi et al. (this issue) described an interaction between speech-gesture synchrony and gesture function. Apparently, communicative head gestures synchronized with the center portions of a verbal backchannel and were temporally aligned with the ends of boundaries. In sum, the exact timing of apexes relative to prominent syllables and lexical affiliates appears to be more complex than suggested by McNeill's synchrony rules, nonetheless, timing seems to have a crucial influence on utterance interpretation.

Another approach to the synchrony between gesture and speech is suggested in that the coordination lies in the process of conceptualization, not necessarily leading to a strict synchrony in production (McNeill, 2005; de Ruiter, 2000a; Kirchhof, 2011). In fact, it has been shown that gestures time shifted by up to 600ms before or after their natural place of occurrence are still semantically integrated by listeners (Kirchhof and de Ruiter, 2012). These figures differ from Leonard and Cummins (2010) results on beat gestures mentioned earlier. It is likely that the reason for the different results lies in the focus put on different gesture types, namely beat (Leonard and Cummins, 2010) vs. deictic and iconic gestures (Kirchhof and de Ruiter, 2012). It is therefore to be suspected that gesture function and timing constraints interact: deictic gestures resemble beat gestures in form, however, in contrast, they do have a clear referential component.

Rather than regard temporal speech-gesture coordination as an intra-speaker phenomenon, researchers have also looked at it as an alignment or entrainment process taking place *between* interlocutors. In this sense, Condon and Ogston (1971) suggested that head movements occur in synchrony with the interlocutor's speech to way bolster communicative attention. Louwerse et al. (2012) used cross-recurrence analysis to characterize behavior matching also called alignment, synchronization, entrainment or temporal convergence between interlocutors in dialogue. They found that head movements and facial behaviors tended to be matched within 1.5 seconds and the relative frequency of recurring states at different lags seemed to reflect the roles of the participants: The instruction follower matches the head movements of the instruction giver more often than the other way round. Also, the interpersonal synchrony in nodding tended to increase over time reflecting the possible social affiliation of the participants. The reasons behind such an adaptive process have often been sought in the low-level process of rhythmic entrainment which enables speakers to closely predict and shadow each other's utterance timing characteristics (Wagner et al., in press).

Feyereisen (1987) noted early that clarification of the computational relations between speech and gesture will depend on the understanding of the precise temporal relationship between them. There is a need to further disambiguate and describe the degrees of synchrony involved in speech-gesture coordination phenomena. Especially, if these phenomena are used to substantiate speech-gesture production models. Paying attention to the methodologies used to investigate synchrony is also essential. A methodological caveat with respect to the interpretation of speech-gesture synchrony comes from Leonard and Cummins (2010, 69), who stated that

> Simultaneity of two events is not evidence that the two events are more tightly coupled (i.e. with less variance) than two events that occur at a fixed lag. Evidence that two events exhibit a functional linkage or coupling must come from examining variability.

We revisit methodological questions with respect to gesture-prosody coordination later in Section 4.3.

4.2. Gesture and prosody similarities, differences, interaction

While communication is clearly the product of multimodal strategies, language is considered as being primarily governed by "combinations of discrete units", organized hierarchically and unfolding linearly (McNeill, 1992), while speech-accompanying gestures are regarded as being mimetic, analog, idiosyncratic and characterized by "continuously varying forms" Goldin-Meadow (1999, 420). Interestingly, gestures have much in common with prosody in their potential for structural nuancing that is non-dis-

crete: in analogy to a speech-accompanying manual gesture, the pitch accent on an utterance can be produced with more or less excursion along an intonation phrase, thus expressing the novelty or importance of the corresponding item to various degrees (Bolinger, 1961; Rietveld and Gussenhoven, 1985; Terken, 1991).

Similar prominence-increasing effects as well as interactions with prosody have been found for gestures (Swerts and Krahmer, 2008) and have been termed *audiovisual prosody*, and are believed to closely interact with the co-occuring speech. The spoken signal and head movements join in with the prominence lending properties of facial expression, such as eyebrow movements (Ekman, 1979; Krahmer and Swerts, 2007; Granström and House, 2007; al Moubayed et al., 2010) to an effect of a multimodal increase in prominence.

The enhancement can serve interactive functions, e.g. by showing a higher degree of attention, and semantic functions, e.g. by signaling information focus (Beskow et al., 2006). Research by Bull and Connelly (1985) suggests that in principle, any body part can assume this highlighting function, and that it is not restricted to hand or head gestures. Audiovisual prosody also appears to facilitate comprehension. Munhall et al. (2004) found that speech amplitude and fundamental frequency correlated with head movement. Additionally, in perceptual experiments with an animated agent the authors showed that word recognition was facilitated by the prominence enhancing contribution of head gestures, along with pitch and intensity. The authors concluded that speech intelligibility varies directly as a function of head motion. Other instances of audiovisual prosody relate to the communication of emotion, attitude, attention, and engagement (Ekman, 1979; Cafaro et al., 2012; Ishii and Nakano, 2008). These have also been found to be expressed prosodically in dimensions such as valency and arousal rather than in discrete categories such as anger, fear or joy (Schröder et al., 2001). Again, mirroring prosody, head gestures have also been found for evaluating or commenting on the ongoing discourse (Poggi et al., 2013).

Karpiński et al. (2009) notice that gesture phrases and intonational phrases bear some structural resemblance: "Both are built around a central prominent event (the accent in speech and the stroke in the gestural modality) and in both cases, this event (is) the only obligatory component of the unit." (Karpiński et al., 2009, 119). Given these similarities, it should come as no surprise that, like gesture, prosody has been linked to very similar interactive functions (cf. Section 1). Some examples are discourse organization, prosodic highlighting (prominence), expression of a speaker's attitude, and structuring the speech into parseable units (Beckman and Pierrehumbert, 1986; Gravano and Hirschberg, 2011; Rietveld and Gussenhoven, 1985; Schröder et al., 2001; Swerts and Geluykens, 1994).

The precise nature of the interaction between audiovisual prosody and verbal prosody, e.g. in prominence highlighting, is hitherto not very well understood. This includes the question of whether both modalities parallel or complement each other, i.e. whether a lack in verbally produced prominence is equalized by gesture or whether the various modalities contribute to prominence in an additive way. Fernández-Baena et al. (this issue) found evidence for an additive effect of both modalities in prominence highlighting. However, temporal co-occurence was less important that a structural resemblance in prominence expression. Interestingly, Ferré (this issue) showed that a marked prosodic structure tends to be reinforced by parallel gesturing, while this is not necessarily the case for marked syntactic constructions. This further strengthens the assumption that there exists a deeper connection between gestural and prosodic expression.

Despite its continuous nature, prosody can express grammaticalized, discrete meanings (Gussenhoven, 1999). At the same time, it is often linked to prosodic meaning universals such as the frequency code (Ohala, 1984) where high tones express smallness and uncertainty, and low tones express large size and assertiveness. It seems likely that co-speech gestures share their gradient and mimetic nature with prosody via *universal metaphors* (Lakoff and Johnson, 1980). However, unlike gesture, prosody and verbal expression share a physiological production mechanism. Hence, they are tightly coupled, as speech will always exhibit some kind of prosody, but the temporal coupling between prosodic events such as stresses and tones and verbal events such as words leaves room for variation.

Modern phonological frameworks such as Autosegmental–Metrical Phonology (Goldsmith, 1990) or Articulatory Phonology (Browman and Goldstein, 1986) model the partial independence existing between prosody and speech, while still regarding both as two layers of a common grammar using a common production and perception system. In Autosegmental–Metrical Phonology, links between two descriptive layers denote temporal co-occurrence, while in Articulatory Phonology, phonological contrasts and structures are explained as the result of temporal coordination between articulatory movements. A similar approach may be suitable to a description of linguistic functions conveyed via a temporally constrained gesture-speech interaction (Bressem and Ladewig, 2011; Jannedy and Mendoza-Denton, 2005; Treffner et al., 2008).

Gesture and prosody share similarities in their non-discreteness, their idiosyncrasy and are probably coordinated motorically. They are linked both temporally, e.g. by some degree of synchrony, and structurally, e.g. by expressive similarity in strength or shape, and can reinforce each other. This idea has received recent support. In a motor task of co-speech tapping, speakers were unable to de-synchronize tapping and emphatic stress, with emphases in either speech or manual behavior automatically lengthening productions in the other domain (Parrell et al., 2011). The strong developmental link has also been supported by Esteve-Gibert and Prieto (this issue), who found tempo-

ral coordination between gesture and speech prosody as early as the babbling phase, with stroke onsets coinciding with onsets of prominent syllables.

With respect to the mechanisms explaining this observed speech-gesture synchrony, Tuite (1993) proposed a rhythmic pulse underlying the production of gesture and originating in the prominence patterns of speech, i.e. the sequences of stressed and unstressed words and syllables. However, there is a great deal of ongoing discussion in the prosody community about the definition, validity and underlying mechanisms of speech rhythm. It can at least be said that if there is a rhythmic pulse underlying or emerging from speech, gestural landmarks very likely correlate with this pulse, given (a) the motoric nature of both processes, (b) the sources of speech and gesture in hand-mouth linkages (cf. Section 4.1), and (c) the similar functions both modalities can convey. One idea that can help explain this coordination stems from a Dynamical Systems perspective. The theory postulates an *entrainment coupling* between various systems on a motor level in cognitive processing (Kelso et al., 1983). This idea has also been taken as a point of departure in the study by Rusiewicz et al. (this issue). Using a perturbation paradigm, the authors examined the coordination of speech, gesture and prosody, and found evidence in favor of a low-level entrainment being at work in speech-gesture coordination.

To summarize, the temporal coordination between prosody and gesture seems stronger than that between gesture and speech in general. Still, the temporal constraints that gesture and verbal expression impose on each other may be less strict or rather different than those between prosody and speech, probably because of their expression in different physiological systems imposing biomechanical constraints and because of the, still unexplained, specific characteristics of the cognitive processes involved in speech and gesture production.

4.3. Measuring temporal coordination

Speech and gesture follow inherently different structural characteristics and are constrained by different physiological systems. As such, the question of how speech and gesture are coordinated in time has no trivial answer. Given the close coordinative affinities between prosody and manual gesture and the similarities between gesture and prosody in their relationship to speech (cf. Section 4.2), several authors have proposed quite early that prosodic landmarks and gestural landmarks should coordinate naturally. First observations concerning prosody and gesture and their correlation were made by Birdwhistell (1952). It was suggested that intonation both in terms of contours and specific points such as pitch accents aligned with gestural movements. Bolinger observed that gestures followed pitch contours up and down, in their main direction of movement (Bolinger, 1983, 1986). McClave (1991) and Loehr (2004, 2012) set out to verify Bolinger's observation on speech-gesture parallelism. The phenomenon occurred occasionally in their data, although they found no significant correlation.

In quantitative analyses on coordination between the modalities, intuitions on speech-gesture synchronization necessitate measurements of anchor points or intervals within the continuous movements constituting gesture and speech (cf. Section 4.1). These points typically correspond to differently estimated *effort maxima*, originally based in early qualitative findings that beat gestures' energy maxima tended to align with prominent syllables (Schegloff, 1984). Several empirical studies examining the timing coordination between various gesture and prosodic anchors are presented in Table 1.

Prominent syllables correspond to the parts of the speech stream which are often assumed to be produced with the most articulatory effort (Eriksson et al., 2001; Tamburini and Wagner, 2007). The timing relations between gesture and prosody investigated in the literature concentrated on pitch accented syllables and different types of gesture effort peaks. The reliance on pitch accents possibly stems from a long-standing consensus that these serve as a major prominence-lending acoustic parameter. The point of maximum effort in the gestural movement in turn is described in the literature with varying degrees of measurement objectivity and with varying definitions of what counts as an observation of maximum effort. Most definitions evoke a kinaesthetic quality of effort or *peak effort* (Kendon, 1972) correlated with abrupt changes in visible movement either as periods of movement acceleration or *strokes* (Kita et al., 1998),[6] as sudden halts or *hits* (Shattuck-Hufnagel et al., 2007), or as maximal movement extensions in space called *apexes* (Leonard and Cummins, 2010). This is paralleled with the muscular and glottal effort in speech as correlated with abrupt acoustic changes in spectral quality and the production of pitch accents (Stetson, 1951; Birdwhistell, 1970).

The gesture stroke has been often associated with effort-based definitions such as the one proposed by Kita et al. (1998, 33): "A phase, in which more force is exerted than neighboring phases, is a stroke. Note that acceleration (and deceleration) is a good indicator of the exerted force". Note that, following McNeill (1992), functionally, the stroke is the most meaningful rather than the most effortful part of a gesture. Additionally, strokes may encompass *stroke holds* with no motion at all. This does not imply that as such, strokes may not correlate with increased expenditure of kinetic energy, due to the intended salience of the meaningful movement phase. Kita et al. (1998, 27) discuss the notion of effort in relation to McNeillian gesture phases (cf. Section 2.1):

> Different types of phases can be identified by different foci of "effort" [...]. In a preparation and a retraction,

[6] See Esteve-Gibert and Prieto (this issue) for a method of identifying strokes through blurring of video frames.

the effort is focused on reaching their respective end points (the beginning of the following stroke and the resting position). In contrast, in a stroke, the effort is focused on "the form of the movement itself – its trajectory, shape, posture" McNeill (1992, p. 376) defines the stroke both on the formal and functional grounds. Functionally, the stroke is the "content-bearing part of the gesture".

Nonetheless, precise measurements of kinetic energy along the duration of different forms of the stroke phase relative to other phases should help guide the association of an essentially functional unit, such as the stroke, with effort in the future.

Beginning with observational studies on the basis of transcribed material, Tuite (1993) found that the strokes of different types of gestures coincided with the nuclear syllable of the affiliated tone group. Karpiński et al. (2009) reported a corpus-based analysis of temporal co-occurrence of gesture phrases, intonational phrases, as well as queries on overlap between two levels of prominences in speech and gesture strokes in Polish. They found evidence for the strict application of the PSR in approx. 40% of cases, and provided several qualitative reasons for the exceptions. These were stroke repetitions, inertial echoes, and fluid hand excursions that often defied the PSR rule. In the case of ill-formed intonational phrases and gestural phrases, the rule was often also not valid and the PSR frequently failed when disrupted by gestural or speech hesitations. Such results led them to suggest a re-definition of the stroke phase in the context of temporal coordination and the PSR. As the word itself seems to suggest, a stroke is a salient and meaningful movement that often ends with a hit. When discrete points within the meaningful movement are found, these discrete points appear to, in turn, align with points of high prominence in speech.

Similar reasoning motivated studies examining hits and apexes of gestural movement. Along with Loehr (2004, 2012), Jannedy and Mendoza-Denton (2005), Yassinik et al. (2004) found that apexes or hits were aligned with pitch accents. There were almost no apexes that did not coincide with a pitch accent in Jannedy and Mendoza-Denton's data. The data reported in Loehr (2012) however, suggest that, again, the term *synchrony* should not be treated as strict co-occurrence, since the distribution of the distances between the apexes and nearest pitch accents exhibited a standard deviation of 341 ms, with the mean lead of the prosodic landmark at half a video frame (17 ms).

In their quest for a more precise speech anchor for gesture, Leonard and Cummins (2010) noted that it is difficult to arrive at atomistic descriptions of gesture with clear reference points within the movement that can be related to similar points in speech. They considered different phonetic candidates as anchors. Among these were the vowel onset of a stressed syllable, the estimated perceptual centre of the syllable and the pitch peak of a stressed syllable. These points were analysed in relation to beat gestures, as their form is less constrained by meaning expression. Reference points in the beat gesture kinematics were measured: movement onset and offset, peak velocities of the extension and retraction phases, as well as the point of maximum extension. Leonard and Cummins (2010) found that the variability for the maximum point of extension was lowest regardless of the speech anchor that it related to. However, similarly to findings by Loehr (2012), the pitch accent on the stressed syllable was most tightly synchronised with the extension peak. Additionally, maximal velocity of the extension phase turned out to align best with the perceptual center and the vowel onset. These results suggest that beat gesture phases are tightly synchronized with prosodic landmarks.

As mentioned earlier, beat gestures tend to have little or no semantic content and their timing and form is relatively unrestricted by meaning expression. Instead, their meaning resides in their "strength and timing", as Leonard and Cummins (2010) note. Contrary to most representational gestures beats are bi-phasic, in that they are produced by a short movement away from the rest state and back (McNeill, 1992; Cassell, 1996), whereas representational gestures typically involve preparation, stroke and retraction phases (cf. Section 2.1), i.e. they are tri-phasic and much slower (Cassell, 1996; Wilson et al., 1996). In sum, beats are simple and unconstrained and therefore, can be used as *kinetic landmarks* for the accompanying speech.

Beats also appear to occur very frequently while speaking. McNeill (1992) found that 44.7% of all gestures produced in a retelling of a Sylvester and Tweety cartoon were beats. In a direction giving corpus, Theune and Brandhorst (2010) found that 32.1% of gestures were beats and reported a high inter-speaker variability in the use of beats relative to other gestures. They also discussed difficulties in distinguishing pure beat gestures from other gestural forms, in particular pointing gestures and iconics. McNeill (1992) proposed a *Beat Filter* listing defining kinematic beat gesture characteristics. Theune and Brandhorst (2010) evaluated the accuracy of the *Beat Filter* by means of agreement measures between annotators using the criteria for classifying beats. They found that this purely physical classification was not highly reliable ($\kappa = .34$). The results suggest that beats are rarely generated as pure beats, e.g. down-and-up sequences of punctual movements with the fist or the palm perpendicular to the floor but are often superimposed on other forms. Their correlation with prosodic landmarks in speech may suggest that prosodic landmarks serve as affordances for the entrainment of punctual movement (Leonard and Cummins, 2010) aiding the functional cohesion of gesture and speech as they unfold in time.

The evidence for simple gestural forms such as manual beats and abrupt stopping points within more complex gesture strokes suggests that these forms closely synchronise with pitch accents, often placed on prominent syllables. Pitch accents signal lexical and sentence stress, and infor-

mation structure in many languages, albeit in a language specific way (Jun, 2007). More research is needed on the differences between superimposed and pure beat coordination with speech. Also, to our knowledge, there are no studies on fluid, continuous gestural movements, and the way in which they are aligned with continuous pitch contours. Rather, the evidence for temporal synchronization stems from landmarks from within what are continuous manual and prosodic movements. Due to their restriction to manual gestures, these studies cannot be extended to other modalities, even though their highlighting function can be taken over by various body parts (Bull and Connelly, 1985).

Similarly to beats and hits, head movements, typically bi-phasic and oscillatory, seem to coordinate well with pitch accented syllables (Fernández-Baena et al., this issue). Hadar et al. (1983) found three modes in communicative head movement frequencies (cf. Section 2.2), slow, ordinary and fast, which they suggested might correlate with prosodic cycles, such as the syllabic and phrasal or *stress kinemes* (Birdwhistell, 1970).

5. Applications, annotations, tools and corpora

In this section, we give an overview of more applied aspects of research on co-speech gesture. We describe how co-speech gestures are annotated (cf. Section 5.1), which software tools (cf. Section 5.2) and corpora (cf. Section 5.3) are available for annotation and analysis, how gesture research has influenced the development of technical systems, and how these systems serve to evaluate existing theories and models (cf. Section 5.4).

5.1. Gesture annotation

Numerous annotation schemata have been developed for the study of gesture, each coming with various advantages, limitations, theoretical assumptions and foci. The choice of an adequate annotation schema is likely to depend on the research question, research field, the intended type of analysis (quantitative or qualitative), the available material and annotation resources and the possible application. A general distinction can be made between annotations focusing more on form (Martell, 2002; Trippel et al., 2004) or on function (Caldognetto and Poggi, 2001; Allwood et al., 2007), while most systems take into account both aspects. Of course, any functional annotation architecture will necessarily make some theoretical assumptions as to the kinds of functions that gestures may have. In the NEUROGES system, the attempt was made to decouple form and function annotation by creating an annotation of form and temporal issues, followed by a functional annotation (Lausberg and Sloetjes, 2009).

Most annotation schemata (cf. Table 2) concentrate on manual gestures and their representational functions. The MUMIN schema is specifically designed to treat pragmatic functions of manual and facial gestures (Allwood et al., 2007). The Multimodal Score (Caldognetto and Poggi, 2001; Caldognetto et al., 2004) is specifically designed for a systematic analysis of hand as well as facial gestures. One aspect where the transcription systems differ significantly is their treatment of gesture segmentation and grouping into phrases. Many systems (Kipp, 2001; Kipp et al., 2007; Lücking et al., 2013; Trippel et al., 2004) implicitly or explicitly follow the segmentation suggestions by McNeill (1992, 2005), later specified by Duncan in her *Annotative Practice*,[7] of separating gesture phrases based on the identification of a stroke, and by identifying preparation and retraction phases prior and after the stroke. A more complex approach towards sequencing these gesture phrases into larger Movement Units was introduced by Kita et al. (1998) and has been taken up and extended in Brugman et al. (2002) and Brugman and Russel (2004). Their ideas have also been taken into account in Kipp's annotation schemata (Kipp, 2001; Kipp et al., 2007) and in Trippel et al. (2004) in a slightly modified fashion by allowing for sequences of single gesture phrases.

All annotation systems focusing on form need to specify ways of doing so. Some make rather vague suggestions and leave the level of descriptive detail to the annotator (Caldognetto and Poggi, 2001; McNeill, 2005; Selting et al., 1998; Selting et al., 2009), while others have decided on very complicated strategies for capturing the gestural details (Martell, 2002). Most systems have tried to compromise between these extremes in order to make the annotations less time consuming, while still capturing potentially relevant detail. Examples of systems trying to find a balance between applicability/usability and information-depth are the SAGA-annotation (Lücking et al., 2013), CoGest (Trippel et al., 2004) and Kipp's system (Kipp et al., 2007). Though differing in the level of detail, all annotations of gesture form take into account aspects of *shape*, *location*, *orientation*, as well as *movement attributes* such as trajectories and the location before and after a dynamic movement. Some systems do not differentiate between single handed and two-handed gestures, while others pay a great deal of attention to this distinction. An example is the NEUROGES system which asks for a specification of the relationship between the hands (Lausberg and Sloetjes, 2009).

Another important aspect refers to the annotation of the speech-gesture interaction. This is usually made explicit through heuristics for annotating the lexical affiliate, e.g. in Kipp et al. (2007). The word co-occurring with the gesture is chosen based on the presence of prosodic prominence, and when unclear, on the annotator's intuition. Caldognetto and Poggi (2001), Caldognetto et al. (2004)

[7] http://mcneilllab.uchicago.edu/analyzing-gesture/intro_to_annotation.html.

Table 2
Key properties of various gesture annotation schemata.

Name/Author(s)	Modality	Functions	Forms
McNeill (2005), Duncan's "annotative practice"	Hands	Functional dimensions (iconicity, metaphoricity, ...)	Qualitative descriptions based on detailed heuristics
Selting et al. (1998, 2009): GAT/ GAT2	Any nonverbal act	Focus on behavior, but iconics, deictics or emblems should be mentioned	Qualitative description of beginning and end, apex (optional)
Kita et al. (1998)	Hands	Syntagmatic rules	Segmentation criteria based on body movement shapes and directions
Kipp (2001)	Hands, arms	Emblems, adaptors, iconics, deictics, metaphorics, beats, phrases and groups	Handedness, dynamic and static features
Kipp et al. (2007)	Hands	Prototypical conversational gesture "lexemes"	Hand trajectory, position, start/end, stroke position, handedness
Lücking et al. (2013)	Hands	Deictic, iconic, discourse gestures; representation techniques: indexing, placing, shaping...	Wrist/palm position, back of hand orientation, movement direction and trajectory
Martell (2002, 2005): FORM	Hands	None	Location, shape, orientation, movement, specified by list of attributes and values per body part
Trippel et al. (2004): CoGest	Hands, arms	None	Gesture source and optional route specification (trajectory and target), location in space, hand shape description with a feature vector similar to FORM
Lausberg and Sloetjes (2009): NEUROGES	Hands	Spatial-functional relations, pointing, space, emphasis, motion, convention, emotion, etc.	Beginning/end, trajectories, locations, handedness
Caldognetto and Poggi (2001) and Caldognetto et al. (2004): Multimodal Score	Hands, head, face	Descriptive typology, paraphrased meaning, meaning typology, semantic gesture-speech coordination (redundant, non-redundant)	Description of perceptual characteristics
Allwood et al. (2007): MUMIN	Hands, head	Feedback, turn organization, sequencing	None

specifically demand a further classification of the gesture's semantic function as redundant, complementary or supplementary to the co-occuring speech. Many annotation systems have co-evolved with well-suited annotation software (cf. Table 3). Table 2 lists the key properties of various annotation schemata.

Table 3
Software tools fitting the needs of co-developed annotation schemata.

Software	Annotation Schema
ELAN	NEUROGES
ANVIL	Kipp's systems, multimodal score
TASX	CoGest
EXMARaLDA	GAT/GAT2

5.2. Annotation tools

Along with annotation schemata, software for gesture annotation and analysis has also been developed (ELAN: Brugman and Russel (2004), TASX: Gut and Milde (2003), EXMARaLDA: Schmidt (2004) or ANVIL: Kipp (2012)). All of these allow for a detailed analysis and annotation of audio and video on various annotation levels or tiers. It is usually possible to import or export annotations to and from related systems, e.g. software targeted to in-depth phonetic analyses such as Praat (Boersma and Weenink, 2008) or according to the needs of database handling. The various tools come with individual specialized capacities, e.g. for motion capture coding extensions with automatic video and audio recognition. Often, annotation systems have been co-developed with software, or software has been specifically designed to meet the needs of particular transcription systems, so that a preference for an annotation system may be a good motivator for choice of software, e.g. by predefined gesture categorizations or the necessary options for linking annotation levels (cf. Table 3). However, most annotation schemata should be flexible enough to work with various annotation tools.

5.3. Multimodal corpora

Due to the technical limitations of video recording and data storage, earlier research on gesture often had to restrict itself to the discussion of sample material or case studies, although the complexity of multimodal corpus construction has already been tackled in a concise way by Mertins et al. (2000). A good overview of existing corpora as well challenges and limits in corpus building can be found in Knight (2011). In recent years, a large number of multimodal corpora have been built according to the needs of different research questions, and along with improved technology such as the availability of large data storage, video tracking and motion capture devices. Of note is that an LREC workshop series has evolved around the topic of multimodal data collections.[8] These work-

[8] http://www.multimodal-corpora.org/.

shops, tackle topics such as annotation schemata and annotation evaluation, as well as data storage and exchange (Kipp et al., 2009). Building a multimodal corpus requires decisions along many dimensions. These include *the number of participants* (monologue, dialogue, multiparty settings), *the number and setting of recording or tracking devices* such as cameras, microphones, motion capture systems or eye trackers and the *conversational setting* (e.g. free conversation vs. task oriented dialogue; friends vs. strangers; same vs. different sex/age/status).

All of these decisions influence the type of the signals (which body parts are shown or not, are the audio signals separated or not) as well as the type of communicative data. In addition, the researcher must keep in mind that the choice of a recording setting will most likely influence the level of invasiveness and, consequently, the ecological validity of the resulting data (Oertel et al., 2013). In corpus building, many potential confounds are often ignored. Subtle details may influence the data, e.g. the use of certain chairs may allow for gesturing more than others, and different room acoustics and atmospheres may lead to different interaction styles. Because of this, it is imperative that researchers pay close attention to the data collection setting so as to avoid unnecessary, albeit controllable, biases.

The multitude of tracking possibilities and recording channels can lead to problems of synchronization between these channels, and also to problems in data reduction and information interpretation. High quality automatic analysis tools for head and motion tracking, and speech recognition, are needed to make large amounts of data manageable, while at the same time, controlling invasiveness. First steps towards solving some of these problems are discussed in Kousidis et al. (2012). Annotating multimodal corpora requires an enormous amount of time. This likely explains the limited number available using the more detailed transcription schemata described in Section 5.1. Because of this, free data collections and initiatives fostering the accessability and interchangeability of corpus data (e.g. AMI,[9] CLARIN-D,[10] HUMAINE[11] or SSPnet[12]) have become extremely valuable to the research community.

5.4. Technical models of co-speech gestures

As gestures have been shown to help speech production, conversation management and perception, a good understanding of their functions and co-production with speech is also essential to building technical applications such as multimodal dialogue models and artificial agents. The impact of adequate head movements in improving comprehension and acceptability of artificial agents has been supported in various studies (Kopp et al., 2008; Beskow et al., 2007; Granström and House, 2007; al Moubayed et al., 2010). The positive impact of manual gestures has also been examined for multimodal dialogue systems and conversational agents (Bergmann et al., 2010; Kipp and Martin, 2009), as well as for humanoid robots (DeSteno et al., 2012; Salem et al., 2012). These studies have reported evidence that gesturing in virtual agents or robots can have decisive effects on perceived lifelikeness, engagement, competence and trustworthiness. Further, gestures made by tutoring agents have been shown to support the learning of new material (Bergmann and Macedonia, 2013; Mayer and DaPra, 2012).

Loosely corresponding to psycholinguistic production models, generating and synthesizing gestural (manual or facial) co-speech behavior is generally conceived of in terms of three major steps to be taken to map a given communicative goal onto graphical behavior animations: (i) communicative content planning, (ii) behavioral realization planning, and (iii) gesture realization. The SAIDA standardization initiative (Kopp et al., 2006; Vilhjálmsson et al., 2007) has set out to formulate XML specification languages for a description of the interfaces between these three stages: FML, the Function Markup Language, and BML, the Behavior Markup Language. As communicative content planning is outside the scope of this overview, it will not be treated in the following paragraphs.

Behavior planning determines behavioral forms that fulfill communicative goals in the current dialogue and discourse context. The solution to this problem is, to some extent, contingent upon the method used for behavior realization. Different approaches to solve this problem have evolved. Most existing systems employ a lexicon-based approach for behavior planning for various types of co-speech gestures (Cassell et al., 2000; Krenn and Pirker, 2004; Poggi, 2001). A more flexible but theoretically more challenging way to generate behaviors is a generative model which has been used for head movements by Heylen et al. (2008).

Similarly, the NUMACK system (Kopp et al., 2004) tried to overcome the limitations of lexicon-based gesture generation by considering patterns of human gesture composition. Based on empirical results, visuo-spatial referent features were linked to morphological gesture features. Another line of research has focused on individual variation. Ruttkay (2007) endowed virtual humans with a unique style in order to appear prototypical of a social or ethnic group. Different styles were, again, defined in a dictionary of meaning-to-gesture mappings with optional modifying parameters to specify the characteristics of a gesture. Hartmann et al. (2006) investigated the modification of gestures to carry a desired expressive content while retaining the original semantics. Bodily expressivity was defined using a small set of dimensions used to modify gesture. These were spatio-temporal extent, fluidity, and power.

[9] www.amiproject.org/.
[10] clarin-d.org/en/.
[11] www.emotion-research.net.
[12] www.sspnet.eu.

An increasingly popular line of research uses data-driven methods to simulate individual speakers gesturing behavior. Stone et al. (2004) recombines motion captured events with new speech samples to recreate coherent multimodal utterances. That way, units of communicative performance are re-arranged while retaining temporal synchrony and communicative coordination that characterizes a person's spontaneous delivery. Neff et al. (2008) used statistical gesture profiles learned from annotated multimodal behavior to generate a character-specific gesture style in a virtual agent that is perceived as more lively and natural. More recently, Chiu and Marsella (2011) used sequential probabilistic models to generate gestural movement from previous motion and audio features. The synthesis by analysis approach was also taken by Sargin et al. (2006, 2008): an automatic audiovisual mapping from speech prosody to head gesture and gesture generation was achieved by a prosody-driven head gesture analysis via a parallel HMM structure. In a study by Yehia et al. (2002), head motion, facial and acoustic features were taken from measurements and a linear regression model was used to generate head motion from the pitch contour.

In an integrated approach to generate iconic gestures, Bergmann and Kopp (2009) combined probabilistic data-driven techniques with rule-based model-driven decision-making within a Bayesian decision network. Being learned from the data of single speakers or a set of speakers, the network was able to predict form features of iconic gestures based on input about the visuo-spatial properties of the referent, the dialog context, or the communicative goal of the speaker. This model was embedded in a more general speech and gesture production architecture. Formulating models in computational terms that integrate speech and gesture production such that they can predict verbal and gestural behavior from given communicative and contextual demands remains a challenging but necessary endeavor: It requires to make mechanisms explicit that often remain implicit and vague in theoretical models due to a lack of solid empirical evidence. By enabling us to test the correctness of their predictions, the implementation of theories may contribute significantly to our understanding of how speech and gestures interact, both covertly within speakers and overtly as part of a natural dialogue.

The approaches to *behavior realization* differ mostly in the motion control algorithms they employ. Systems that emphasize naturalness of motion usually employ *motion capturing* to record movements performed by human actors and simply replay them as a primitive form of motion control. Using this technique, the range of motions that can be produced is limited by the range of the stored data. It has found its main application for producing behaviors whose form is predefined by the combination and adjustment of a limited set of stereotypical motions, or which need only a very limited adaptation, such as breathing or postural sway (Neff et al., 2008; Stone et al., 2004). More flexible animation is possible by motion control algorithms that perform some form of online control over the movement, using techniques like key-framing, inverse kinematics, or interpolation (Hartmann et al., 2006). The highest flexibility and generative power is provided by procedural animation. Using this approach, motion control is exerted online and throughout the entire movement by specialized controllers that can employ an explicit model of the target trajectory or some other model of the flow of control parameters over time. This technique is often used for the generation of behaviors with very specific external features such as iconic gestures (Bergmann and Kopp, 2009). Recent approaches employ sequential probabilistic models that treat movement as a time-varying probability distribution in some control parameter space, often employing latent variables to account for inter-feature covariation (Lee and Marsella, 2006). In general, the different approaches apply differently well to individual behaviors, and need to be combined in order to synthesize convincing co-speech gesturing.

Automatic recognition of manual gesture can be seen as a processing chain that begins with capturing the position and orientation of hand movements together with the angles of the finger joints for determining the hand posture. Overall, this is achieved either by *contact methods* using data gloves or markers for a tracking system mounted on the user's hands and arms, or *non-contact methods* based on video or infrared cameras (Wu and Huang, 1999). These methods rely on the fitting of an assumed hand or body model to the input data, which subsequently allows for reading out specific angles or position vectors. These body models can be either *kinematic* and based on the skeletal structure of joints and limbs or *dynamic*, describing motion as the result of the application of forces and torques. Body models are helpful, as they provide for an estimation of missing sensor information, a prediction of movement trajectories, an adaptation to individual user characteristics, and detection or rejection of impossible or implausible gestural configurations or body movements. Just like in speech, gesture recognition can furthermore be seen as a segmentation problem, referring to the problem of filtering out a gesture's expressive phase and determining units of meaning from the continuous stream of data delivered by the sensing devices. The segmentation is either performed separately as a preprocessing step, or built in as an integral part of the recognition model as, for example, in approaches using Hidden Markov Models (HMM). These exploit the spatial configuration of the arms as feature vectors and may rely on difference images of video streams. That way, Eickeler et al. (1998) achieve recognition rates of about 93% for 24 isolated gestures. Combinations of HMM and ANN techniques have also been used successfully to recognize five different dynamic gestures from monocular video images (Corradini, 2002). If relying on preprocessing, the gesture's semantics is ignored and explicit spatio-temporal cues are exploited to determine unit borders such as sign changes of the first derivative (Wexelblat, 1995), local minima of hand tension (Harling and

Edwards, 1997), or rule based groupings of whole-hand configurations or *gestlets* (Spoons et al., 1993).

For head gestures, most systems concentrate on the recognition of nods and shakes, only a few involve other gesture types. Recognition rates are often very good with nod and shake recognition up to 100% (Morency et al., 2005), more complex head movement types are recognized with 87.3% accuracy (Akakin and Sankur, 2011). A thorough review of head gesture recognition systems can be found in Bousmalis et al. (2012).

6. Concluding remarks

Despite the recently increased interest in gesture and speech interaction, we remain far from understanding its precise nature. Still, we expect that our future understanding will profit from the, now widespread, availability of technical tools such as annotation software and affordable solutions for building multimodal corpora. It should nevertheless be kept in mind that the mere collection of data is of little relevance without meaningful ways of interpretation and analysis. These rely on available models and theories of gesture shapes, movements, functions and processing. Likewise, theories remain of little impact, if they cannot be adequately tested. Quantifiable operationalizations of terms and concepts, but also integrated implementations of gesture and speech processing in robots or virtual agents provide valuable environments for the formal and empirical evaluation of our models and theories. Targeting a wide audience, we hope to have provided a comprehensive overview on the topic. For the future, we hope for a strengthened interdisciplinary dialogue on speech and gesture interaction for the benefit of our entire research community.

Acknowledgments

We warmly thank Mingyuan Chu, Maciej Karpiń ski, Spyros Kousidis, Bernd Möbius and Marc Swerts for valuable and constructive feedback on an earlier version of this paper. Department Digital (Bielefeld) kindly provided help with the preparation of Fig. 1. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Collaborative Research Center 673 “Alignment in Communication”.

References

Abercrombie, David, 1954. Gesture. ELT Journal 9 (1), 3–12.

Akakin, Hatice Çinar, Sankur, Bülent, 2011. Robust classification of face and head gestures in video. Image Vision Computing 29 (7), 470–483.

al Moubayed, S., Beskow, J., Granström, B., 2010. Auditory visual prominence: from intelligence to behavior. Journal on Multimodal User Interfaces 3 (4), 299–309.

Alibali, Martha Wagner, DiRusso, Alyssa A., 1999. The function of gesture in learning to count: more than keeping track. Cognitive Development 14 (1), 37–56.

Alibali, M.W., Heath, D.C., Myers, H.J., 2001. Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. Journal of Memory and Language 44, 169–188.

Alibali, M.W., Kita, S., Young, A.J., 2000. Gesture and the process of speech production: we think, therefore we gesture. Language and Cognitive Processes 15, 593–613.

Allwood, Jens, Loredana, Cerrato, Jokinen, Kristiina, Navarretta, Constanza, Paggio, Patrizia, 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation 41, 273–287.

Allwood, Jens, Cerrato, Loredana, 2003. A study of gestural feedback expressions. First Nordic Symposium on Multimodal Communication. Copenhagen, Denmark, pp. 7–22.

Altorfer, Andreas, Jossen, Stefan, Würmle, Othmar, Käsermann, Marie-Louise, Foppa, Klaus, Zimmermann, Heinrich, 2000. Measurement and meaning of head movements in everyday face-to-face communicative interaction. Behaviour Research Methods Instruments and Computers 32 (1), 17–32.

Bavelas, Janet B., 1994. Gestures as part of speech: methodological implications. Research on Language and Social Interaction 27, 201–221.

Bavelas, J., Gerwing, J., Sutton, C., Prevost, D., 2008. Gesturing on the telephone: independent effects of dialogue and visibility. Journal of Memory and Language 58, 495–520.

Bavelas, Janet B., Coates, Linda, Johnson, Trudy, 2006. Listener responses as a collaborative process: the role of gaze. Journal of Communication 52 (3), 566–580.

Becker, Raymond, Cienki, Alan, Bennett, Austin, Cudina, Christina, Debras, Camille, Fleischer, Zuzanna, Haaheim, Michael, Müller, Torsten, Stec, Kashmiri, Zarcone, Alessandra. 2011. Aktionsarten, speech and gesture. In Proceedings of GESPIN2011: Gesture and Speech in Interaction, Bielefeld, Germany.

Beckman, Mary, Pierrehumbert, Janet, 1986. Intonational structure in Japanese and English. Phonology Yearbook 3, 255–309.

Bergmann, Kirsten, Kopp, Stefan, 2006. Verbal or visual? how information is distributed across speech and gesture in spatial dialogue. In: Schlangen, David, Fernandez, R. (Eds.), Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue. Universitätsverlag, Potsdam, Germany, pp. 90–97.

Bergmann, Kirsten, Kopp, Stefan, 2009. Increasing expressiveness for virtual agents – autonomous generation of speech and gesture for spatial description tasks. In: Decker, K., Sichman, J., Sierra, G., Castelfranchi, C. (Eds.), Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009). Ann Arbor, MI, pp. 361–368, IFAAMAS.

Bergmann, Kirsten, Kopp, Stefan, 2010. Systematicity and idiosyncrasy in iconic gesture use: empirical analysis and computational modeling. Gesture in Embodied Communication and Human–Computer Interaction. Springer, Berlin, pp. 182–194.

Bergmann, Kirsten, Macedonia, Manuela, 2013. A virtual agent as vocabulary trainer: Iconic gestures help to improve learners' memory performance. In: Aylett, R., Krenn, B., Pelachaud, C., Shimodaira, H. (Eds.), Intelligent Virtual Agents, . In: Lecture Notes in Artificial Intelligence. Springer, Berlin, Heidelberg, pp. 139–148.

Bergmann, Kirsten, Kopp, Stefan, Eyssel, Friederike, 2010. Individualized gesturing outperforms average gesturing – evaluating gesture production in virtual humans. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (Eds.), Proceedings of the 10th International Conference on Intelligent Virtual Agents (LNCS 6356). Springer, Berlin, Heidelberg, pp. 104–117.

Bergmann, Kirsten, Aksu, Volkan, Kopp, Stefan. 2011. The relation of speech and gestures: temporal synchrony follows semantic synchrony. In: Proceedings of GESPIN2011: Gesture and Speech in Interaction, Bielefeld, Germany.

Bergmann, Kirsten, Kahl, Sebastian, Kopp, Stefan. 2013. Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In: Proceedings of the International Conference on Intelligent Virtual Agents (IVA 2013).

Beskow, J., Granström, B., House, D. 2006. Focal accent and facial movements in expressive speech Fonetic 2006, Working Papers, pp. 52:9–52:12.

Beskow, J., Granström, B., House, D., 2007. Analysis and synthesis of multimodal verbal and non-verbal interaction for animated interface agents. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (Eds.), Verbal and Nonverbal Communication Behaviours. Springer, Berlin, pp. 250–263.

Bevacqua, Elisabetta. 2009. Computational Model of Listener Behavior for Embodied Conversational Agents. Ph.D Thesis, Université Paris 8, Paris, France.

Birdwhistell, Ray L., 1952. Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture. University of Michigan Press.

Birdwhistell, Ray L., 1970. Kinesics and Context. In: Essays on Body Motion Communication. University of Pennsylvania Press, Philadelphia, PA.

Boersma, Paul, Weenink, David. 2008. Praat: Doing phonetics by computer (version 5.0.13). <http://www.praat.org>, March 18.

Bolinger, Dwight, 1961. Contrastive accent and contrastive stress. Language 37, 87–96.

Bolinger, Dwight, 1982. Intonation in nondeclaratives. Chicago Linguistics Society Parasession on Nondeclaratives. Chicago Linguistics Society, Chicago, pp. 1–15.

Bolinger, Dwight, 1983. Intonation and gesture. American Speech 58, 156–174.

Bolinger, Dwight, 1986. Intonation and Its Parts: Melody in Spoken English. Stanford University Press.

Bousmalis, Konstantinos, Mehu, Marc, Pantic, Maja, 2012. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: a survey of related cues databases and tools. Image and Vision Computing 31 (2), 203–221.

Bressem, Jana, Ladewig, Silva, 2011. Rethinking gesture phases: articulatory features of gestural movement. Semiotica 184 (1/4), 53–91.

Browman, Catherine, Goldstein, Louis., 1986. Towards an articulatory phonology. Phonology Yearbook 3, 219–252.

Brugman, Hennie, Russel, Albert. 2004. Annotating multimedia/multimodal resources with elan. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, pp. 2065–2068. <http://tla.mpi.nl/tools/tla-tools/elan/>.

Brugman, Hennie, Wittenburg, Peter, Levinson, Stephen C., Kita, Sotaro. 2002. Multimodal annotations in gesture and sign language studies. In: Third International Conference on Language Resources and Evaluation, pp. 176–182.

Bull, Peter, Connelly, Gerry, 1985. Body movement and emphasis in speech. Journal of Nonverbal Behavior 9 (3), 169–187.

Buss, Samuel R., 2003. 3D Computer Graphics: A Mathematical Introduction with Open GL. Cambridge University Press.

Butterworth, B., Beattie, G., 1978. Gesture and silence as indicators of planning in speech. In: Campbell, R.N., Smith, P.T. (Eds.), Recent Advances in the Psychology of Language: Formal and Experimental Approaches. Plenum, New York.

Butterworth, B., Hadar, U., 1989. Gesture, speech, and computational stages: A reply to mcneill. Psychological Review 96, 168–174.

Cafaro, Angelo., Vilhjálmsson, Högni Hannes, Bickmore, Timothy W., Heylen, Dirk, Johannsdottir, Kamilla R., Valgardsson, Gunnar Steinn. 2012. First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In: IVA, pp. 67–80.

Caldognetto, Emanuela M., Poggi, Isabella, Cosi, Piero, Cavicchio, Federica, Merola, G. 2004. Multimodal score: an ANVIL based annotation scheme for multimodal audio-video analysis. In: Martin, J.-C., Os, E.D., Kühnlein, P., Boves, L., Paggio, P., Catizone, R. (Eds.), Proceedings of the LREC-Workshop on Multimodal Corpora, Lisbon, Portugal, pp. 29–33.

Caldognetto, Magno, Poggi, Isabella. 2001. Dall'analisi della multimodalità quotidiana alla costruzione di agenti animati con facce parlanti ed espressive. In: Cosi, P., Caldognetto, M. (Eds.), Multimodalità e multimedialità della communicatione, Atti delle XI Giornate di Studio del G.F.S. Unipress, Padova.

Cassell, Justine, 1996. A framework for gesture generation and interpretation. In: Cipolla, R., Pentland, A. (Eds.), Computer Vision in Human–Machine Interaction. Cambridge University Press.

Cassell, Justine, Sullivan, Joseph, Prevost, Scott, Churchill, Elizabeth F. (Eds.), 2000. Embodied Conversational Agents. MIT Press.

Cerrato, Loredana. 2007. Investigating Communicative Feedback Phenomena across Languages and Modalities. Ph.D Thesis, KTH Computer Science and Communication, Department of Speech, Music and Hearing, Stockholm, Sweden.

Chiu, Chung-Cheng, Marsella, Stacy, 2011. How to train your avatar a data driven approach to gesture generation. In: The 11th International Conference on Intelligent Virtual Agents (IVA), Taipei, Taiwan, May 2–6.

Christenfeld, Nicholas, Schachter, Stanley, Bilous, Frances, 1991. Filled pauses and gestures: it's not coincidence. Journal of Psycholinguistic Research 20 (1), 1–10.

Chui, K., 2005. Temporal patterning of speech and iconic gestures in conversational discourse. Journal of Pragmatics 37, 871–887.

Cienki, Alan, Müller, Cornelia, 2008. Metaphor, gesture, and thought. In: Gibbs, R.W., Jr.Jr. (Ed.), The Cambridge Handbook of Metaphor and Thought. Cambridge University Press, pp. 483–502.

Clark, Herbert H., Wilkes-Gibbs, Deanna, 1986. Referring as a collaborative process. Cognition 22, 1–39.

Clark, Herbert H., Brennan, Susan, 1991. Grounding in communication. In: Lockhead, G., Pomerantz, J. (Eds.), Perspectives on Socially Shared Cognition. APA Books, Washington, pp. 127–149.

Condon, W.S., Ogston, W.D., 1971. Speech and body motion synchrony of the speaker-hearer. In: Horton, D.L., Jenkins, J.J. (Eds.), Perception of Language. Merrill, Columbus, Ohio.

Corradini, A., 2002. Real-time gesture recognition by means of hybrid recognizers. In: Wachsmuth, Ipke, Sowa, Timo (Eds.), Gesture and Sign Language in Human–Computer Interaction. Springer, Berlin, Heidelberg, New York, pp. 34–46.

de Ruiter, J.P., Bangerter, A., Dings, P., 2012. Interplay between gesture and speech in the production of referring expressions: investigating the tradeoff hypothesis. Topics in Cognitive Science 4 (2), 232–248.

de Ruiter, J.P., 2000a. The production of gesture and speech. In: McNeill, David (Ed.), Language and Gesture. Cambridge University Press, pp. 248–311.

de Ruiter, J.P., 2000b. The production of gesture and speech. In: McNeill, David (Ed.), Language and Gesture. Cambridge University Press, Cambridge, UK, pp. 284–311.

de Ruiter, J.P., Noordzij, M.L., Newman-Norlund, S., Newman-Norlund, R., Hagoort, P., Levinson, S.C., Toni, I., 2010. Exploring the cognitive infrastructure of communication. Interaction Studies 11 (1), 51–77.

DeSteno, D., Breazeal, C., Frank, R.H., Pizarro, D., Baumann, J., Dickens, L., Lee, J., 2012. Detecting the trustworthiness of novel partners in economic exchange. Psychological Science 23, 1549–1556.

Dittmann, A.T., Llewellyn, L.G., 1968. Relationship between vocalizations and head nods as listener responses. Journal of Personality and Social Psychology 9, 79.

Dittmann, A.T., Llewelyn, L.G., 1969. Body movement and speech rhythm in social conversation. Journal of Personality and Social Psychology 23, 283–292.

Dobrogaev, S.M., 1929. Ucnenie o reflekse v problemakh iasykovedeniia (observations on reflexes and issues in language study). Iazykovedenie i Materializm, 105–173.

Duncan, Starkey, 1972. Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23, 283–292.

Eickeler, Stefan, Kosmala, Andreas, Rigoll, Gerhard. 1998. Hidden Markov model based continuous online gesture recognition. In:

Proceedings of the International Conference on Pattern Recognition (ICPR), Brisbane, Australia, August 1998, pp. 1206–1208.
Ekman, P., 1979. About brows: emotional and conversational signals. In: von Cranach, M., Foppa, K., Lepenies, W., Ploog, D. (Eds.), Human Ethology: Claims and Limits of a New Discipline. Cambridge University Press, Cambridge, pp. 169–202.
Ekman, P., Friesen, W.V., 1972. Hand movements. Journal of Communication 22, 353–374.
Eriksson, A., Thunberg, G., Traunmüller, H. 2001. Syllable prominence: a matter of vocal effort, phonetic distinctness and top-down processing. In: Proceedings of EUROSPEECH, Aalborg, Denmark, pp. 399–402.
Esteve-Gibert, Núria, Prieto, Pilar. Infants temporally coordinate gesture-speech combinations before they produce their first words. Speech Communication, this issue.
Fernández-Baena, Adso, Montaño, Raúl, Antonijoan, Marc, Roversi, Arturo, Miralles, David, Alías, Francesc. Gesture synthesis adapted to speech emphasis. Speech Communication, this issue.
Ferré, G. 2010. Timing relationships between speech and co-verbal gestures in spontaneous French. In: Language Resources and Evaluation Conference (LREC). Workshop on Multimodal Corpora, Malta.
Ferré, Gaëlle. A multimodal approach to markedness in spoken French. Speech Communication, this issue.
Feyereisen, P., 1987. Gestures and speech, interactions and separations: a reply to McNeill. Psychological Review 94, 168–174.
Gentilucci, Maurizio, Dalla Volta, Riccardo, 2007. The motor system and the relationship between speech and gesture. Gesture 7 (2), 159–177.
Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., Wagner, S., 2001. Explaining math: gesture lightens the load. Psychological Science 12 (6), 516–522.
Goldin-Meadow, Susan, 1999. The role of gesture in communication and thinking. Trends in Cognitive Sciences 3 (11), 419–429.
Goldin-Meadow, Susan, 2003. Beyond words: the importance of gesture to researchers and learners. Child Development 71 (1), 231–239.
Goldin-Meadow, Susan, Alibali, M.W., Church, S.W., 1993. Transitions in concept acquisition: using the hand to read the mind. Psychological Review 100, 279–297.
Goldsmith, John, 1990. Autosegmental and Metrical Phonology. Blackwell, Oxford.
Goodwin, Charles, 1981. Conversational organization: interaction between speakers and hearers. Academic Press, New York.
Granström, Björn, House, David. 2007. Inside out – acoustic and visual aspects of verbal and non-verbal communication. In Proceedings of the XVIth International Congress of the Phonetic Sciences, Saarbrücken, Germany, pp. 11–18.
Gravano, Agustin, Hirschberg, Julia, 2011. Turn-taking cues in task-oriented dialogue. Computer, Speech and Language 25 (3), 601–634.
Gussenhoven, Carlos, 1999. Discreteness and gradience in intonational contrasts. Language and Speech 42 (2–3), 283–305.
Gut, Ulrike, Milde, Jan Thorsten, 2003. Annotation and analysis of conversational gestures in the TASX environment. KI 17, 34–49.
Habets, B., Kita, S., Shao, Z., Özyürek, A., Hagoort, P., 2011. The role of synchrony and ambiguity in speech-gesture integration during comprehension. Journal of Cognitive Neuroscience 23 (8), 1845–1854.
Hadar, U., Steiner, T.J., Grant, E.C., Clifford Rose, F., 1983. Kinematics of head movements accompanying speech during conversation. Human Movement Science 2, 35–46.
Hadar, Uri, Steiner, T.J., Grant, E.C., Clifford Rose, F., 1984. The timing of shifts of head postures during conversation. Human Movement Science 3, 237–245.
Hadar, Uri, Steiner, T.J., Rose, Clifford F., 1985. Head movement during listening turns in conversation. Journal of Nonverbal Behavior 9, 214–228.
Harling, P., Edwards, A., 1997. Hand tension as a gesture segmentation cue. In: Harling, P., Edwards, A. (Eds.), Progress in Gestural Interaction: Proceedings of the Gesture Workshop 1996. Springer, Berlin, Heidelberg, New York, pp. 75–87.
Harrison, S. 2013. The temporal coordination of negation gestures in relation to speech. In: Proceedings of TiGeR 2013, Tilburg, NL.
Hartmann, Björn, Mancini, Maurizio, Pelachaud, Catherine, 2006. Implementing expressive gesture synthesis for embodied conversational agents. In: Gibet, Sylvie, Courty, Nicolas, Kamp, Jean-François (Eds.), Gesture in Human–Computer Interaction and Simulation. Springer, Berlin, Heidelberg, pp. 188–199.
Heldner, Mattias, Hjalmarsson, Anna, Edlund, Jens, 2012. Continuer relevance spaces. In: Asu, Eva-Liina, Lippus, Pärtel (Eds.), Nordic Prosody. Proceedings of the XIth Conference, Tartu 2012. Peter Lang, Frankfurt, pp. 137–146.
Heylen, Dirk. 2005. Challenges ahead: head movements and other social acts in conversations. In: Proceedings of AISB 2005, pp. 45–52.
Heylen, Dirk, 2006. Head gestures, gaze and the principles of conversational structure. International Journal of Humanoid Robotics 3 (03), 241–267.
Heylen, Dirk, 2008. Listening heads. In: Wachsmuth, Ipke, Knoblich, Guenther (Eds.), Modeling Communication with Robots and Virtual Humans. Springer, Berlin, pp. 241–259.
Heylen, Dirk, Bevacqua, Elisabetta, Pelachaud, Catherine, Poggi, Isabella, Gratch, Jonathan, Schröder, Marc, 2011. Generating listening behaviour. In: Petta, Paolo, Pelachaud, Catherine, Cowie, Roddy (Eds.), Emotion-Oriented Systems: The Humaine Handbook. Springer, Berlin, pp. 321–348.
Hoetjes, Marieke, Krahmer, Emiel, Swerts, Marc. Does our speech change when we cannot gesture? Speech Communication, this issue.
Holler, J., Stevens, R., 2009. The effect of common ground on how speakers use gesture. Journal of Language and Social Psychology 26, 4–27.
Holler, Judith, Beattie, Geoffrey, 2003. Pragmatic aspects of representational gestures. Gesture 3 (2), 127–154.
Holler, Judith, Stevens, Rachel, 2007. The effect of common ground on how speakers use gesture and speech to represent size information. Journal of Language and Social Psychology 26 (4), 4–27.
Hostetter, A.B., Alibali, M.W., 2007. Raise your hand if you're spatial: relations between verbal and spatial skills and gesture production. Gesture 7 (1), 73–95.
Hostetter, A.B., Potthoff, A.L., 2012. Effects of personality and social situation on representational gesture production. Gesture 12 (1), 62–83.
Hostetter, Autumn B., Alibali, Martha Wagner, 2008. Visible embodiment: gestures as simulated action. Psychonomic Bulletin and Review 15, 495–514.
Hostetter, Autumn B., Alibali, Martha Wagner, Kita, Sotaro, 2007. I see it in my hand's eye: representational gestures reflect conceptual demands. Language and Cognitive Processes 22 (3), 313–336.
Ishi, Carlos Toshinori, Ishiguro, Hiroshi, Hagita, Norihiro, Analysis of relationship between head motion events and speech in dialogue conversation. Speech Communication, this issue.
Ishii, Ryo, Nakano, Yikiko I., 2008. Estimating user's conversational engagement based on gaze behaviors. Intelligent Virtual Agents. In: Lecture Notes in Computer Science, vol. 5208. Springer, pp. 200–207.
Iverson, J.M., Thelen, E., 1999. Hand, mouth and brain – the dynamic emergence of speech and gesture. Journal of Consciousness Studies 6 (11–12), 19–40.
Iverson, J.M., Goldin-Meadow, S., 1998. Why do people gesture as they speak? Nature 396, 228.
Jakobson, Roman, 1972. Motor signs for 'yes' and 'no'. Language and Society 1, 91–96.
Jannedy, S., Mendoza-Denton, N., 2005. Structuring information through gesture and intonation. Interdisciplinary Studies on Information Structure 3, 199–244.
Jokinen, Kristiina, Nishida, Masfumi, Yamamoto, Seiichi. 2010. On eye-gaze and turn-taking. In: Proceedings of EGIHMI'10, New York, USA, pp. 118–123.
Jun, Sun-Ah, 2007. Prosodic typology. In: Jun, Sun-Ah (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, New York, pp. 430–462, chapter 16.
Karpiński, Maciej, Jarmołowicz-Nowikow, Ewa, Malisz, Zofia, 2009. Aspects of gestural and prosodic structure of multimodal utterances in

Polish task-oriented dialogues. Speech and Language Technology 11, 113–122.
Kelso, J.A.S., Tuller, B., Harris, K.S., 1983. A dynamic pattern perspective on the control and coordination of movement. In: McNeilage, P.F. (Ed.), The Production of Speech. Springer-Verlag, New York, pp. 138–173.
Kendon, Adam, 1972. Some relationships between body motion and speech. In: Siegman, A., Pope, B. (Eds.), Studies in Dyadic Communication. Pergamon, New York, pp. 177–210.
Kendon, Adam, 1980. Gesture and speech: two aspects of the process of utterance. In: Key, Mary R. (Ed.), Nonverbal Communication and Language. Mouton, The Hague, pp. 207–227.
Kendon, Adam, 2003. Some uses of the head shake. Gesture 2, 147–182.
Kendon, Adam, 2004. Gesture – Visible Action as Utterance. Cambridge University Press.
Kim, Jeesun, Cvejic, Erin, Davis, Christopher, Tracking eyebrows and head gestures associated with spoken prosody. Speech Communication, this issue.
Kipp, Michael. 2001. From human gesture to synthetic action. In: Proceedings of the Workshop on "Multimodal Communication and Context in Embodied Agents" held in conjunction with the Fifth International Conference on Autonomous Agents (AGENTS), Montreal, Canada, May.
Kipp, Michael. 2004. Gesture generation by imitation: From human behavior to computer character animation. Ph.D Thesis, Saarland University.
Kipp, Michael, 2012. Multimodal annotation, querying and analysis in ANVIL. In: Maybury, Mark T. (Ed.), Multimedia information extraction, John Wiley and Sons Inc, Hoboken, NJ, pp. 531–368.
Kipp, Michael, Martin, J.-C., 2009. Gesture and emotion: can basic gestural form features discriminate emotions? Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII-09). IEEE Press.
Kipp, Michael, Martin, Jean-Claude, Paggio, Patrizia, Heylen, Dirk (Eds.), 2009. Multimodal Corpora: From Models of Natural Interaction to Systems and Applications. Lecture Notes in Computer Science, vol. 5509. Springer, Berlin, Heidelberg.
Kipp, Michael, Neff, Michael, Albrecht, Irene, 2007. An annotation scheme for conversational gestures: how to economically capture timing and form. Language Resources and Evaluation 41, 325–339.
Kirchhof, Carolin. 2011. So what's your affiliation with gesture? In: Proceedings of GESPIN2011: Gesture and Speech in Interaction, Bielefeld, Germany.
Kirchhof, Carolin, de Ruiter, J.P. 2012. On the audiovisual integration of speech and gesture. In: Fifth Conference of the International Society for Gesture Studies, Lund, Sweden.
Kita, S., Davies, T.S., 2009. Competing conceptual representations trigger co-speech representational gestures. Language and Cognitive Processes 24 (5), 795–804.
Kita, Sotaro, 2000. How representational gestures help speaking. In: McNeill, David (Ed.), Language and Gesture. Cambridge University Press, pp. 162–185.
Kita, Sotaro, Özyürek, Asli, 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language 48 (1), 16–32.
Kita, Sotaro, Gijn, Ingeborg van, Hulst, Harry van der, 1998. Movement phases in signs and co-speech gestures* and their transcription by human coders. In: Wachsmuth, Ipke, Fröhlich, Martin (Eds.), Gesture and Sign Language in Human–Computer Interaction, . In: Lecture Notes in Artificial Intelligence, vol. 1371. Springer, pp. 23–35.
Knight, Dawn, 2011. The future of multimodal corpora. RBLA 2, 391–415.
Kopp, Stefan, Allwood, Jens, Grammar, Karl, Ahlsén, Elisabeth, Stocksmeier, Thorsten, 2008. Modeling embodied feedback with virtual humans. In: Wachsmuth, Ipke, Knoblich, Günther (Eds.), Modeling Communication with Robots and Virtual Humans. Springer-Verlag, Berlin, pp. 18–37.
Kopp, Stefan, Bergmann, Kirsten, Kahl, Sebastian. 2013. A spreading-activation model of the semantic coordination of speech and gesture. In: Proceedings of the 35th Annual Meeting of the Cognitive Science Society (COGSCI 2013).
Kopp, Stefan, Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K., Vilhjálmsson, H., 2006. Towards a common framework for multimodal generation in ecas: the behavior markup language. Proceedings of Intelligent Virtual Agents (IVA). In: Gratch, J. (Ed.), . In: LNAI, vol. 4133. Springer-Verlag, pp. 205–217.
Kopp, Stefan, Tuller, B., Cassell, Justine. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In: Sharma, R., Darrell, T. (Eds.), Proceedings of the International Conference on Multimodal Interfaces (ICMI'04), pp. 97–104.
Kousidis, Spyros, Malisz, Zofia, Wagner, Petra, Schlangen, David. 2013. Exploring annotation of head gesture forms in spontaneous human interaction. In: TiGeR 2013, Tilburg Gesture Research Meeting.
Kousidis, Spyros, Pfeiffer, Thies, Malisz, Zofia, Wagner, Petra, Schlangen, David. 2012. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In: Proceedings of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue, Skamania Lodge, Stevenson, WA.
Krahmer, Emiel, Swerts, Marc, 2007. The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. Journal of Memory and Language 57 (3), 396–414.
Krauss, R.M., Hadar, U., 1999. The role of speech-related arm/hand gestures in word retrieval. In: Campbell, R., Messing, L. (Eds.), Gesture, Speech, and Sign. Oxford University Press, pp. 93–116.
Krenn, Brigitte, Pirker, Hannes. 2004. Defining the gesticon: language and gesture coordination for interacting embodied agents. In: Proceedings of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters, Leeds, UK, March 29–April 1, pp. 107–115.
Kunin, Mikhail, Osaki, Yasuhiro, Cohen, Bernard, Raphan, Theodore, 2007. Rotation axes of the head during positioning, head shaking and locomotion. Journal of Neurophysiology 98 (5), 3095–3108.
Lakoff, John, Johnson, Mark, 1980. Methaphors We Live by. The University of Chicago Press, Chicago.
Lausberg, Hedda, Sloetjes, Han, 2009. Coding gestural behaviour with the NEUROGE-ELAN system. Behaviour Research Methods, Instruments, and Computers 41 (3), 841–949.
Lee, Jina, Marsella, Stacy. 2006. Nonverbal behavior generator for embodied conversational agents. In: Proceedings of Sixth International Conference on Intelligent Virtual Agents (IVA), Marine del Rey, USA, August 21–23, pp. 243–255.
Leonard, Thomas, Cummins, Fred. 2009. Temporal alignment of gesture and speech. In: Jarmołowicz-Nowikow, Ewa, Juszczyk, Konrad, Malisz, Zofia, Szczyszek, Michał (Eds.), Proceedings of GESPIN2009: Gesture and Speech in Interaction, Poznań, Poland, pp. 1–6.
Leonard, Thomas, Cummins, Fred, 2010. The temporal relation between beat gestures and speech. Language and Cognitive Processes 26 (10), 1457–1471.
Levelt, W.J.M., Richardson, G., Hey, W.L., 1985. Pointing and voicing in deictic expressions. Journal of Memory and Language 24, 133–164.
Loehr, D. 2004. Gesture and Intonation. Ph.D Thesis, Georgetown University, Washington, D.C.
Loehr, Daniel, 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology. Journal of the Association for Laboratory Phonology 3, 71–89.
Loevenbruck, H., Dohen, M., Vialin, C., 2009. Pointing is 'special'. In: Fuchs, S., Loevenbruck, H., Pape, D., Perrier, P. (Eds.), Some Aspects of Speech and the Brain. Peter Lang, New York, pp. 211–258.
Louwerse, Max M., Dale, Rick, Bard, Ellen G., Jeuniaux, Patrick, 2012. Behaviour matching in multimodal communication is synchronized. Cognitive Science 36 (8), 1404–1426.
Lücking, Andy, Bergmann, Kirsten, Hahn, Florian, Kopp, Stefan, Rieser, Hannes, 2013. Data-based analysis of speech and gesture: the Bielefeld

speech and gesture alignment corpus (SaGA) and its applications. Journal of Multimodal User Interfaces 7 (1–2), 5–18.
Martell, C., 2002. Form: An extensible, kinematically based gesture annotation scheme. In: Proceedings of ICSLP'02, pp. 353–356.
Martell, Craig. H., 2005. Form – an extensible, kinematically based gesture annotation scheme. In: van Kuppevelt, J.C.J. (Ed.), Advances in Natural Multimodal Dialogue Systems. Springer, pp. 79–96.
Mayer, R.E., DaPra, C.S., 2012. An embodiment effect in computer-based learning with animated pedagogical agents. Journal of Experimental Psychology: Applied 18, 239–252.
McClave, E. 1991. Intonation and Gesture. Ph.D Thesis, Georgetown University.
McClave, Evelyn, 1994. Gestural beats: the rhythm hypothesis. Journal of Psycholinguistic Research 23 (1), 45–66.
McClave, Evelyn, 2000. Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32, 855–878.
McNeill, David, 1985. So you think gestures are nonverbal? Psychological Review 92, 350–371.
McNeill, David, 1987. So you do think gestures are nonverbal? reply to Feyereisen (1987). Psychological Review 92, 350–371.
McNeill, David, 1989. A straight path to where? reply to Butterworth and Hadar. Psychological Review 94, 499–504.
McNeill, David, 1992. Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago.
McNeill, David, 2005. Gesture and Thought. University of Chicago Press.
Mertins, Inge, Moore, Roger, Gibbon, Dafydd (Eds.), 2000. Handbook of Multimodal and Spoken Dialogue Systems: Resources and Product Evaluation. Kluwer Academic Publishers, Norwell.
Morency, Louis-Philippe, Sidner, Candace, Lee, Christopher, Darrell, Trevor. 2005. Contextual recognition of head gestures. In: Proceedings of the Seventh International Conference on Multimodal Interfaces, pp. 18–24.
Morrel-Samuels, Palmer, Krauss, Robert M., 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Human Learning and Memory 18 (3), 615–622.
Munhall, Kevin G., Jones, Jeffery A., Callan, Daniel E., Kuratate, Takaaki, Vatikiotis-Bateson, Eric, 2004. Visual prosody and speech intelligibility. Psychological Science 15 (2), 133–137.
Neff, Michael, Kipp, Michael, Albrecht, Irene, Seidel, Hans-Peter, 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics 27 (1), 5.1–5.2.
Nobe, Shuichi, 2000. Where to most spontaneous representational gestures actually occur with respect to speech? In: McNeill, David (Ed.), Language and Gesture. Cambridge University Press, pp. 186–198.
Oertel, Catharine, Cummins, Fred, Edlund, Jens, Wagner, Petra, Campbell, Nick, 2013. D64 – a corpus of richly recorded conversational interaction. Journal of Multimodal User Interfaces 7 (1–2), 19–28.
Ohala, John, 1984. An ethological perspective on common cross-language utilization of f0 of voice. Phonetica 41, 1–16.
Özyürek, A., Willems, R.M., Kita, S., Hagoort, P., 2007. On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. Journal of Cognitive Neuroscience 19 (4), 605–616.
Parrell, B., Goldstein, L., Lee, S., Byrd, D. 2011. Temporal coupling between speech and manual motor actions. In: Proceedings of the Ninth International Seminar on Speech Production.
Poggi, I., D'Errico, F., Vincze, L. 2010. Types of nods. The polysemy of a social signal. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, pp. 17–23.
Poggi, Isabella. 2001. The lexicon and the alphabet of gesture, gaze, and touch. In: Proceedings of Intelligent Virtual Agents – Third International Workshop (IVA 2001), pp. 235–236, Madrid, Spain, September 10–11.
Poggi, Isabella, D'Errico, Francesca, Vincze, Laura, 2013. Comments by words, face and body. Journal of Multimodal User Interfaces 7, 67–78.
Priesters, M.A., Mittelberg, I. 2013. Individual differences in speakers' gesture spaces: multi-angle views from a motion-capture study. In: Proceedings of the Tilburg Gesture Research Meeting (TiGeR), June 19–21.
Rietveld, Toni, Gussenhoven, Carlos, 1985. On the relation between pitch excursion size and prominence. Journal of Phonetics 13, 299–308.
Rochet-Capellan, A., Laboissière, Rafael, Galván, Arturo, Schwartz, Jean-Luc, 2008. The speech focus position effect on jaw-finger coordination in a pointing task. Journal of Speech, Language and Hearing Research 51 (6), 1507–1521.
Rosenfeld, H.M., Hancks, M., 1980. The nonverbal context of verbal listener responses. In: Key, Mary Ritchie (Ed.), The Relationship of Verbal and Nonverbal Communication. Mouton Publishers, The Hague, The Netherlands, pp. 193–206.
Roth, Wolff-Michael, 2001. Gestures: their role in teaching and learning. Review of Educational Research 71 (3), 365–392.
Roustan, Benjamin, Dohen, Marion. 2010. Co-production of contrastive focus and manual gestures: temporal coordination and effects on the acoustic and articulatory correlates of focus. In: Speech Prosody 2010, Chicago, IL.
Rowbotham, Samantha J., Holler, Judith, Lloyd, Donna, Wearden, Alison. The semantic interplay of speech and co-speech gestures in the description of pain sensations. Speech Communication, this issue.
Rusiewicz, H.L., Shaiman, S., Iverson, J., Szuminsky, N. Effects of perturbation and prosody on the coordination of speech and gesture. Speech Communication, this issue.
Ruttkay, Zsófia, 2007. Presenting in style by virtual humans. COST 2102 Workshop (Vietri). In: Esposito, Anna, Faúndez-Zanuy, Marcos, Keller, Eric, Marinaro, Maria (Eds.), . In: Lecture Notes in Computer Science, vol. 4775. Springer, ISBN 978-3-540-76441-0, pp. 23–36.
Salem, Maha, Kopp, Stefan, Wachsmuth, Ipke, Rohlfing, Katharina, Joublin, Frank, 2012. Generation and evaluation of communicative robot gesture. International Journal of Social Robotics, 1875-4805, 201–217, Special Issue on Expectations, Intentions, and Actions.
Sargin, Mehmet, Aran, Oya, Karpov, Alexey, Ofli, Ferda, Yasinnik, Yelena, Wilson, Stephen, Erzin, Engin, Yemez, Yucel, Tekalp, Murat A. 2006. Combined gesture-speech analysis and speech driven gesture synthesis. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Toronto, Canada, pp. 893–896.
Sargin, Mehmet E, Yemez, Yucel, Erzin, Engin, Tekalp, Ahmet M, 2008. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (8), 1330–1345.
Schegloff, E.A., 1984. On some gestures' relation to talk. In: Atkinson, J.M., Heritage, J. (Eds.), Structures of Social Action. Cambridge University Press, Cambridge, pp. 266–298.
Schmidt, T. 2004. Transcribing and annotating spoken language with EXMARaLDA. In: Proceedings of the LREC-Workshop on XML-based Richly Annotated Corpora, Lisbon. European Language Resources Association (ELRA).
Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S. 2001. Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proceedings of EUROSPEECH, vol. 1, pp. 87–90.
Selting, Margret, Auer, Peter, Barden, Birgit, Bergmann, Jörg, Couper-Kuhlen, Slizabeth, Günthner, Susanne, Quasthoff, Uta, Meier, Christoph, Schlobinski, Peter, Uhmann, Susanne, 1998. Gesprächsanalytisches Transkriptionssystem (GAT). Linguistische Berichte 173, 91–122.
Selting, Margret, Auer, Peter, Bergmann, Jörg, Bergmann, Pia, Birkner, Karin, Couper-Kuhlen, Elizabeth, Deppermann, Arnulf, Gilles, Peter, Günthner, Susanne, Hartung, Martin, Kern, Friederike, Mertzlufft, Christine, Meyer, Christian, Morek, Miriam, Oberzaucher, Frank, Peters, Jörg, Quasthoff, Uta, Schütte, Wilfried, Stukenbrock, Anja, Uhmann, Susanne, 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT2). Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, 353–402.
Shattuck-Hufnagel, Stefanie, Yasinnik, Yelena, Veilleux, Nanette, Renwick, Margaret, 2007. A method for studying the time alignment of

gestures and prosody in American English: 'hits' and pitch accents in academic-lecture-style speech. In: Esposito, Anna, Bratanić, Maja, Keller, Eric, Marinaro, Maria (Eds.), Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue, . In: NATO Security Through Science series E: Human and Societal Dynamics, vol. 18. IOS Press, Washington, DC.

Slobin, Dan I., 1996. From "thought and language" to thinking for speaking. In: Gumperz, J.J., Levinson, S.C. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, pp. 70–96.

So, W.C., Kita, S., Goldin-Meadow, S., 2009. Using the hands to identify who does what to whom: gesture and speech go hand-in-hand. Cognitive Science 33, 115–125.

Spoons, D., Sparrell, C., Thorisson, K., 1993. Integrating simultaneous input from speech, gaze and hand gestures. In: Maybury, M. (Ed.), Intelligent Multimedia Interfaces. AAAI Press/MIT Press, Cambridge, pp. 257–276.

Stetson, Raymond Herbert, 1951. Motor Phonetics. North-Holland, Amsterdam.

Stone, Matthew, DeCarlo, Doug, Oh, Insuk, Rodriguez, Christian, Stere, Adrian, Lees, Alyssa, Bregler, Chris, 2004. Speaking with hands: creating animated conversational characters from recordings of human performance. ACM Transactions on Graphics 23 (3), 506–513.

Streeck, Jürgen, 2008. Depicting by gesture. Gesture 8 (3), 285–301.

Swerts, M., Geluykens, R., 1994. Prosody as a marker of information flow in spoken discourse. Language and Speech 37, 21–43.

Swerts, Marc, Krahmer, Emiel, 2008. Facial expressions and prosodic prominence: comparing modalities and facial areas. Journal of Phonetics 36 (2), 219–238.

Tamburini, Fabio, Wagner, Petra. 2007. On automatic prominence detection for German. In: Proceedings of Interspeech 2007, Antwerp, Belgium, pp. 1809–1812.

Terken, Jacques, 1991. Fundamental frequency and perceived prominence. Journal of the Acoustical Society of America 89 (4), 1768–1776.

Theune, Mariët, Brandhorst, Chris, 2010. To beat or not to beat: beat gestures in direction giving. In: Kopp, Stefan, Wachsmuth, Ipke (Eds.), Gesture in Embodied Communication and Human–Computer Interaction, . In: Lecture Notes in Artificial Intelligence, vol. 5934. Springer, pp. 195–206.

Tomlison, R.D., Cheung, R., Blakeman, A., 2000. Naso-occipital vestibulo-ocular reflex responses in normal subjects. IEEE Engineering in Medicine and Biology Magazine 19 (2), 43–47.

Treffner, Paul, Peter, Mira, Kleidon, Mark, 2008. Gestures and phases: the dynamics of speech-hand communication. Ecological Psychology 20, 32–64.

Trippel, Thorsten, Gibbon, Dafydd, Thies, Alexandra, Milde, Jan-Torsten, Looks, Karin, Hell, Benjamin, Gut, Ulrike. 2004. Cogest: a formal transcription system for conversational gesture. In: Proceedings of LREC 2004, Lisbon, Portugal.

Truong, Khiet P., Poppe, Ronald, de Kok, Iwan, Heylen, Dirk. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In: Proceedings of INTERSPEECH 2011, Florence, Italy. International Speech Communication Association, pp. 2973–2976.

Tuite, Kevin, 1993. The production of gesture. Semiotica 1/2, 83–105.

Urban, Christian. 2011. Temporales Alignment von Sprache und Gestik, B.Sc. Thesis, Bielefeld University.

Vendler, Zeno, 1967. Linguistics in Philosophy. Cornell University Press, Ithaca, NY.

Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, Stefan, 2007. The behavior markup language: recent developments and challenges. Proceedings of Intelligent Virtual Agents (IVA). In: Lecture Notes in Artificial Intelligence, vol. 4722. Springer, pp. 99–111.

Wagner, Petra, Inden, Benjamin, Malisz, Zofia, Wachsmuth, Ipke, in press. Interaction phonology. In: Wachsmuth, Ipke, Jaecks, Petra, de Ruiter, J.P. (Eds.), Towards a New Theory of Communication. Benjamins.

Wexelblat, Alan, 1995. An approach to natural gesture in virtual environments. ACM Transactions on Computer–Human Interaction (TOCHI) 2 (3), 179–200.

Wilson, Andrew D., Bobick, Aaron F., Cassell, Justine, 1996. Recovering the temporal structure of natural gesture. International Conference on Automatic Face and Gesture Recognition. IEEComputer Society Press, Los Alamitos, CA.

Włodarczak, Marcin, Buschmeier, Hendrik, Malisz, Zofia, Kopp, Stefan, Wagner, Petra. 2012. Listener head gestures and verbal feedback expressions in a distraction task. In: Proceedings of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue, Skamania Lodge, Stevenson, WA.

Wu, Ying, Huang, Thomas S., 1999. Vision-based gesture recognition: a review. In: Braffort, Annelies, Gherbi, Rachid, Gibet, Sylvie, Teil, Daniel, Richardson, James (Eds.), Gesture-Based Communication in Human–Computer Interaction, . In: Lecture Notes in Computer Science, Vol. 1739. Springer, Berlin, pp. 103–115.

Yassinik, Y., Renwick, M., Shattuck-Hufnagel, S., 2004. The timing of speech-accompanying gesture with respect to prosody. Proceedings of the International Conference: From Sound to Sense. MIT, Cambridge, pp. C97–C102, June 10–13.

Yehia, C., Kuratate, Takaaki, Vatikiotis-Bateson, Eric, 2002. Linking facial animation, head motion and speech acoustics. Journal of Phonetics 30 (3), 555–568.

Yoganandan, Narayan, Pintar, Frank A, Zhang, Jiangyue, Baisden, Jamie L, 2009. Physical properties of the human head: mass, center of gravity and moment of inertia. Journal of Biomechanics 42 (9), 1177–1192.

Further reading

Cerrato, Loredana, Svanfeldt, Gunilla, 2005. A method for the detection of communicative head nods in expressive speech. In: Allwood, J., Dorriots, B., Nicholson, S. (Eds.), Gothenburg Papers in Theoretical Linguistics 92: Proceedings from The Second Nordic Conference on Multimodal Communication. Göteborg University, Sweden, pp. 153–165.

Erdem, Ugur Murat, Sclaroff, Stan, 2002. Automatic detection of relevant head gestures in American Sign Language communication. Proceedings of the International Conference on Pattern Recognition. IEEE Computer Society Press, pp. 10460–10463.

Kapoor, Ashish, Picard, Rosalind W., 2001. A real time nod and shake detector. Proceedings of the 2001 Workshop on Perceptive User Interfaces. ACM, pp. 1–5.

Li, Renxiang, Taskiran, Cüneyt M., Danielsen, Mike. 2007. Head pose tracking and gesture detection using block motion vectors on mobile devices. In: Mobility '07: Proceedings of the Fourth International Conference on Mobile Technology, Applications and Systems and the First International Symposium on Computer Human Interaction in Mobile Technology, pp. 572–575.

Lu, Peng, Huang, Xiangsheng, Zhu, Xinshan, Wang, Yangsheng, 2005. Head gesture recognition based on Bayesian network. Proceedings of the Second Iberian Conference on Pattern Recognition and Image Analysis. In: Lecture Notes in Computer Science. Springer-Verlag, pp. 495–499.

Nguyen, Laurent, Odobez, Jean-Marc, Gatica-Perez, Daniel, 2012. Using self-context for multimodal detection of head nods in face-to-face interactions. Proceedings of the International Conference on Multimodal Interaction. ACM, pp. 289–292.

Petra Wagner
ZofiaMalisz
StefanKopp
E-mail address: zofia.malisz@uni-bielefeld.de