On the temporal domain of co-speech gestures: syllable, phrase or talk spurt?

David House, Simon Alexanderson and Jonas Beskow
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

Abstract

This study explores the use of automatic methods to detect and extract hand gesture movement co-occuring with speech. Two spontaneous dyadic dialogues were analyzed using 3D motion-capture techniques to track hand movement. Automatic speech/non-speech detection was performed on the dialogues resulting in a series of connected talk spurts for each speaker. Temporal synchrony of onset and offset of gesture and speech was studied between the automatic hand gesture tracking and talk spurts, and compared to an earlier study of head nods and syllable synchronization. The results indicated onset synchronization between head nods and the syllable in the short temporal domain and between the onset of longer gesture units and the talk spurt in a more extended temporal domain.

Introduction

There is currently considerable interest in the interaction between speech and gesture, and in particular the temporal relationship between prosody and gesture (Wagner et al., 2014). Kendon (1980) followed by McNeill (1992) have divided gestures into basic types differing in temporal scope. These are gesture units, gesture phrases and gesture phases. The gesture unit, the longest temporal domain, is the interval of gestural movement bounded by a period of non-movement. A gesture unit is comprised of one or more gesture phrases each of which can be divided up into a sequence of gesture phases. The stroke phase of a gesture phrase is particularly interesting in terms of prosody. Some of these strokes (also called "beat" gestures) often coincide and appear to be synchronized with prosodic and intonational peaks related to prominence such as pitch accents. For example, in studies by Leonard and Cummings (2011) and Loehr (2012), a correlation was found between the apices of strokes and focal accents in intonation. Flecha-Garcia (2007) studied alignment between eyebrow movement and pitch accents, and synchronization between three different phases of head nods and stressed syllables carrying focal accent was found by Alexanderson et al. (2013a, 2013b).

Synchronization between the phrase level of intonation and gesture phrases has also been studied by e.g. Karpinski et al. (2009) and Loehr (2012). Karpinski et al. (2009) found less synchronization between intonational phrases and gesture phrases, but they did note a high degree of overlap and a general "centering tendency" for semantically related gesture phrases and major intonational phrases. Loehr (2012) found a higher degree of synchronization between the gesture phrase and the intermediate intonational phrase as defined by Beckman and Pierrehumbert (1986).

Studying synchronization between speech and gesture has played an important role in building theories of human communication which approach speech and gesture production as arising from a common generation process (Kendon, 2004; McNeill, 2005). However, as seen above, most of the synchronization has been found in shorter domains of the accented syllable and the stroke gesture phase. In this study, we aim to investigate the relationship between longer domains of the gesture unit and talk spurts automatically extracted from spontaneous dialogue in order to test the hypothesis that there is not only a relationship between the intermediate phrase and gesture as detailed by Loehr (2012) but also between a longer period of speaker activation (the talk spurt) and the gesture unit. We further compare the timing results from our earlier studies of head nods and syllables with the timing relationship found between the gesture unit and the talk spurt. Ultimately we wish to explore the correspondences between co-speech, gestural domains and prosodic domains to try and more closely define the temporal domains of co-speech gestures.

Method

Corpus description

Portions of two dialogues taken from the Spontal corpus of Swedish dialogue were used for this investigation. The database, containing more than 60 hours of unrestricted conversation in over 120 dialogues between pairs of speakers, is comprised of synchronized high-quality audio and video recordings (high definition) and motion capture for body and head movements for all recordings with a frame rate of 100 frames per second (Edlund et al. 2010). During the recordings, the participants were seated in a sound studio and allowed to speak about any topic of their choice for 30 minutes. They remained in a seated position throughout the recording session. Figure 1 shows two video frames taken from the corpus. In this study we used a randomly selected five-minute passage from each of the two dialogues: Dialogue 1 between a male and a female participant, and Dialogue 2 between two male participants.

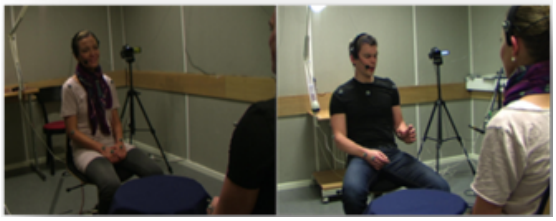

Figure 1. Frames taken from the spontal corpus of spontaneous dialogue

Hand movement detection

The motion data consist of the 3D positions of the motion capture markers attached to each subject. The total marker set contained 12 markers placed on the upper body and is illustrated in Figure 2.

Figure 2. Motion capture markers with 12 markers per subject.

Two different methods were used to quantify hand movements from the data and to divide the extracted movement into discrete sequences. Both methods are based on the conditions in the data corpus whereby the two participants are seated, facing each other, with their hands at a resting position on their laps. The first method (Method 1) is a simple, naïve method and interprets vertical displacement above a certain fixed threshold as gesture movement. Each gesture unit is defined as the time period from one resting position to the next.

The second method (Method 2) uses the velocity of the hand markers as the basis for segmentation. Typically, gesture units exhibit continuous hand motion except for short periods of time, when for example the hands change direction in a stroke. Our algorithm first detects periods of motion larger than 0.001 m/s and then merges segments with separation in time less than 200 ms. During a second pass, segments shorter than 200 ms are removed as well as segments which are performed in a horizontal plane (defined by a margin of 2.5 cm vertical distance). The last filter was introduced to remove non-communicative movement such as fidgeting.

Manual annotation

Two annotators independently annotated gesture phrases and phases using the ELAN annotation tool (Sloetjes and Wittenburg, 2008). The annotation procedure was carried out in two steps. The first step was to watch the dialogue at full speed and mark the beginning and end of any communicative hand gesture, one speaker at a time. The second step consisted of watching the gestures in slow motion and segmenting each gesture phrase into gesture phases following McNeill (1992) and Kita et al. (1998). The specified phases were the following: preparation, stroke, retraction, and pre- and post-stroke holds. All phases were optional except stroke which was an obligatory element of each marked gesture.

Automatic speech activity detection

To determine the speech activity at each frame, a voice activity detection algorithm (Laskowski et al., 2004) was applied to the audio recordings from the near microphones attached to the subjects. The speech activity was then divided into intervals of continuous speech, or talk spurts, following Brady (1968) and Heldner et

al. (2011). In this process, silent intervals of less than 200 ms are not regarded as silence but are integrated into the talk spurt. The process resulted in two sets of talk spurt segments for each dialogue, one for each subject. This procedure also captures short utterances such as humming and feedback signals which were not relevant to this study. Therefore, we introduced a pre-processing stage removing segments shorter than 500 ms.

Results

Evaluation of gesture extraction methods

To evaluate the automatic extraction methods we compared the results of the two methods with the manual annotations on a frame-by-frame basis for presence vs. absence of gestures for each dialogue. These results can then be compared to the average frame-by-frame agreement between the two annotators as an upper baseline (see Table 1). Accuracy is higher for method 2 than for method 1 for both dialogues. Therefore, results obtained from method 2 are used in the following analyses.

Table 1. Annotator agreement and automatic gesture extraction accuracy for the two dialogues and the two methods

	Annotator agreement	Accuracy: method 1	Accuracy: method 2
Dialogue 1	96%	80%	87%
Dialogue 2	99%	88%	96%

In cases where there were erroneous segments derived from method 2, a manual analysis of the characteristics of these segments was carried out. False positive gesture segments were in large caused mainly by pose shifts and fidgeting. Some undetected gestures (false negatives) were present when gestures were only performed by the fingers. Also long stroke holds were not detected as part of gestures as they do not exhibit any motion. However, as the motion capture data has a high frame rate, the method generally was more precise in detecting the exact onset and offset of the hand motion. This is of general importance for investigation of timing aspects and synchrony and can potentially be used to resolve inter-annotator non-agreement concerning the gesture segment boundaries.

Co-occurrence of gesture units and talk spurts

The number of talk spurts co-occurring with the automatically extracted gesture units varied considerably between the subjects and the dialogues. Dialogue 1 could be characterized as a dialogue rich in gesture with gesture units co-occurring in well over half of all the talk spurts, while dialogue 2 contained much fewer gesture sequences represented in only about 15% of the talk spurts. A general trend was that the talk spurts tended either to coincide temporally with a gesture unit throughout most of the duration of the talk spurt, or else the talk spurt contained no gestures at all.

Synchronization of talk spurts and gestures

The temporal relationships between the automatically extracted gesture units and the talk spurts are presented as box and whisker plots in Figure 3. The plots were calculated by measuring the timing difference between the onset of the talk spurt and the onset of the gesture unit. Positive values indicate that the talk spurt leads the gesture unit in time. There is considerable variation in the relationship between onset times, but there is a central tendency in which the onset of the talk spurt slightly precedes the onset of the gesture unit. Greater variability is seen for the two speakers with the most gestures in Dialogue 1.

Synchronization of head nods and syllables

In an earlier study (Alexanderson et al., 2013a) temporal synchronization was studied between head nods annotated as having a beat function and anchor points of the syllable. Both the head nods and the syllable anchor points were automatically extracted. Figure 4 shows the time difference between two different anchor-points of the syllable (onset and nucleus) and three different phases of the nod: peak velocity of the downward phase (p1), max rotation (p2) and peak velocity of the upward phase (p3) for one speaker from the spontal corpus. The timing relationship between the gesture and the syllable does not seem to be influenced by the choice of syllable anchor-point. On an average the nod begins slightly before the syllable onset with the maximum rotation of the nod centering on the syllable nucleus.

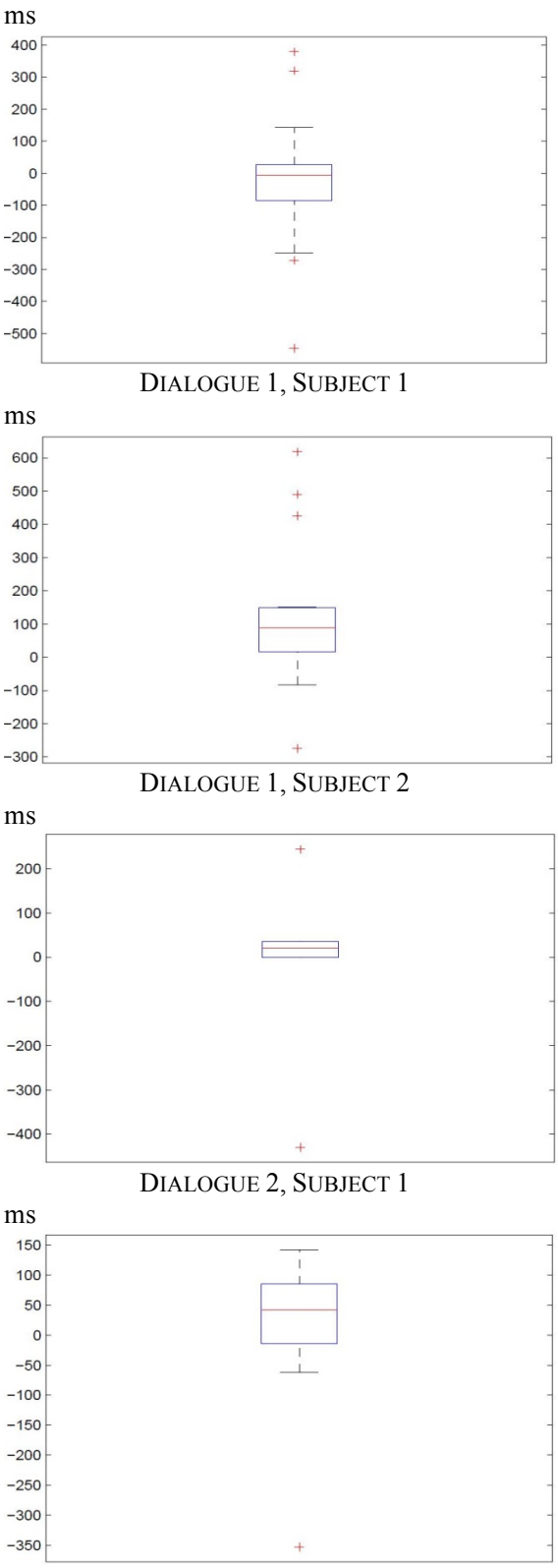

Figure 3. Box and whisker plots showing the temporal relationship between gesture unit and talk spurt. Positive values indicate that the talk spurt precedes the gesture unit.

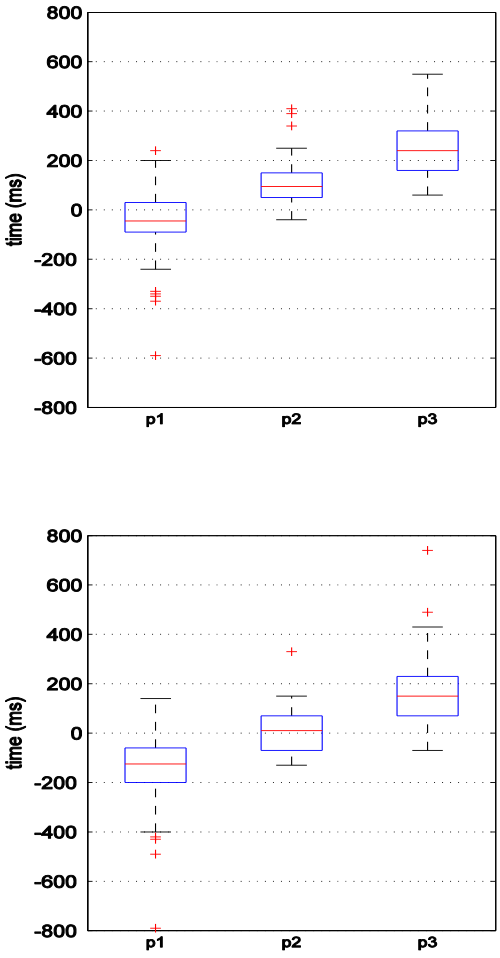

Figure 4. Timing of syllable anchor-points, onset (top) and nucleus (bottom), with respect to three different phases of the head nod: peak velocity of the downward phase (p1), max rotation (p2) and peak velocity of the upward phase (p3). Negative values indicate that the head nod phase precedes the syllable anchor point (from Alexanderson et al. 2013a).

Discussion

In this study we have explored the use of automatic methods to locate co-speech gestures and the co-occurring speech in longer sequences of spontaneous dialogue. We have investigated the temporal aspects of these sequences and seen a co-occurrence of what has in the literature been termed gesture units with automatically extracted talk spurts. This study was limited to two dialogues, but even in only two dialogues, we see a large variation in the incidence of gestures occurring during the talk spurts. One of the dialogues was rich in gesture with over half

of the talk spurts co-occurring with gesture. The other dialogue showed much less gesture activity for both participants. This points to the optionality of producing gesture units which temporally coincide with a complete talk spurt. However, in the gesture-rich dialogue, the gesture units were often found to coincide with the talk spurt and together formed a coordinated speech-gesture unit.

The use of an automatic analysis of the temporal synchronization of gesture movement and talk spurts can help us in our understanding of the domain of co-occurring speech and gesture. The results obtained here are consistent with the findings by Loehr (2012) and Karpinski et al. (2009) in that there is less absolute synchrony between gesture and speech on the phrase level than on the syllable level. The gesture sequences extracted by the automatic gesture movement analysis presented here most closely correspond to the gesture units rather than gesture phrases. The synchronization between the gesture units and the talk spurts shown in Figure 3 displays considerable variation, but also shows a central tendency where the onset of gesture tends to be coordinated with the onset of the talk spurt with the talk spurt slightly preceding the onset of the gesture unit on an average. The initiation phase of the talk spurt may be preceded by gaze activity and posture shifts, but hand movement seems be initiated slightly after the beginning of speech.

This constitutes a timing trend contrary to that appearing between head motion and the syllable as shown in Figure 4. Beat gestures can thereby be seen to share the time domain of the syllable while gesture units share the time domain of the talk spurt. Moreover, this could indicate that on a global temporal domain, speech precedes gesture, while on the local domain of the syllable where beat gestures and accented syllables co-occur, gesture precedes the prosodic correlate having the same function.

Finally, the analysis of gesture and speech in a longer temporal domain points to the possibility of defining the talk spurt as analyzed here as a speech correlate to the gesture unit. This domain is longer than the intonational phrase and can be seen as bodily activation in both speech and gesture comprising an important temporal domain in spontaneous dialogue.

Acknowledgements

The work reported here has been funded by the Bank of Sweden Tercentenary Foundation (P12-0634:1) and the Swedish Research Council (VR 2010-4646). We would also like to thank Meg Zellers for help in preparing and executing the gesture annotation and to Jens Edlund for assistance with the Spontal corpus.

References

Alexanderson S, House D and Beskow J (2013a). Extracting and analyzing head movements accompanying spontaneous dialogue. In *Proc. Tilburg Gesture Research Meeting.* Tilburg University, The Netherlands.

Alexanderson S, House D and Beskow J (2013b). Aspects of co-occurring syllables and head nods in spontaneous dialogue. In *Proc. of 12th International Conference on Auditory-Visual Speech Processing (AVSP2013).* Annecy, France.

Beckman M and Pierrehumbert J (1986). Intonational structure in Japanese and English. *Phonology Yearbook III*: 15-70.

Brady P T (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal,* 47: 73-91.

Edlund J, Beskow J, Elenius K, Hellmer K, Strömbergsson S and House D (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari N, Choukri K, Maegaard B, Mariani J, Odjik J, Piperidis S, Rosner M and Tapias D, eds, *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valetta, Malta, 2992-2995.

Flecha-Garcia M L (2007). Non-verbal communication in dialogue: Alignment between eyebrow raises and pitch accents in English. In *Proceedings of CogSci-2007*. Austin, Texas, USA, 1753.

Heldner M, Edlund J, Hjalmarsson A and Laskowski K (2011). Very short utterances and timing in turn-taking. In *Proceedings of Interspeech 2011*. Florence, Italy, 2837-2840.

Karpinskyi M, Jarmolowicz-Nowikow E and Malisz Z (2009). Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. *Speech and language technology,* 11: 113-122.

Kendon A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key M R ed, *The relationship of verbal and nonverbal communication*. The Hague: Mouton, 207-227.

Kendon A (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.

Kita S, van Gijn I and van der Hulst H (1998). Gesture and Sign Language in Human-Computer Interaction. In Wachsmuth I and Frölich M eds, *Lecture Notes in Computer Science* 1371: Springer, 23-35.

Laskowski K, Jin Q and Schultz T (2004). Crosscorrelation-based multispeaker speech activity detection. Proceedings of *Interspeech 2004*. Jeju Island, South Korea.

Leonard T and Cummins F (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes,* 26: 1457−1471.

Loehr D (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology. *Journal of the Association for Laboratory Phonology*, 3: 71-889.

McNeill D (1992). *Hand and mind: What gestures reveal about thought*. Chicago: The University of Chicago Press.

McNeill D (2005). *Gesture and thought.* Chicago: University of Chicago Press.

Sloetjes H and Wittenburg P (2008). Annotation by category – ELAN and ISO DCR. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Wagner P, Malisz Z and Kopp S (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57: 209-232.