Top-down-Bottom-up Experiments on Detecting Co-speech Gesturing in Conversations

Kristiina Jokinen
Institute of Computer Science
University of Tartu
Kristiina.jokinen@ut.ee

Abstract

Automatic analysis of conversational videos and detection of gesturing and body movement of the partners is one of the areas where technology development has been rapid. This paper deals with the application of video techniques to human communication studies, and focuses on detecting communicative gesturing in conversational videos. The paper sets to investigate the *top-down-bottom-up* methodology, which aims to combine the two approaches used in interaction studies: the human annotation of the data and the automatic analysis of the data.

1 Introduction

Conversations form a social system whereby the interlocutors exchange information about their intentions, interests, and feelings. The participants use verbal and non-verbal means to give feedback and construct common understanding with their partner. Non-verbal communication (Argyle 1988) has been long studied focusing especially on gesturing (Kendon 2004), gaze (Argyle and Cook 1976), and various prosodic and paralinguistic issues (Schuller and Batliner 2013). However, it is only recently that advanced technology has given an opportunity to automatically detect these signals in such a robust way that also interaction studies can benefit of the objective views and of the automatic detection of signals; we talk about Social Signal Processing, which refers to data-directed statistical and machine-learning studies of the verbal and non-verbal signals exchanged in communication. Social signals indicate interest, emotions, affect, etc. and include a wide range of various behavioural signals like gesturing, gaze, laughing, coughing etc.

However, social signal processing requires large data-sets for enabling machine-learning studies and usually also golden standard corpora, or annotated corpora which provide reference point for the evaluation of the algorithms and models. Given the huge work and resource requirements for manual annotation, various algorithms and tools have been developed to assist in the initial analysis of the data, or conduct the segmentation automatically.

In this context, our studies also deploy novel technology in human communication studies, and explore the *top-down-bottom-up* methodological approach and its use in social signal processing. The aim is to provide an objective basis for human annotations concerning conversational partners' head, hand, and body movements, while also taking into account the interpretation of the events in their conversational space.

In particular, the paper focusses on video analysis and gesture recognition technology that enables observations of the human speakers and their movement on video recordings. The recognition technology, described in more detail in Vels and Jokinen (2015), decomposes the observed movement into three gesturing parts (body, head and feet), and regards them as separate activities. In this paper, this technology is used in human interaction studies, and the recognized gesturing is visualized together with the participants' speech, so as to correlate conversational participants' movements with their speaking and listening activity.

The paper is structured as follows. After a short introduction to the methodology in Section 2 and gesturing in Section 3, the paper presents the video processing technology and the data used in the experiments in Section 4. Results are discussed in Section 5, and conclusions and future work in Section 6.

2 *Top-down and bottom-up* Methodology

2.1 Top-down-bottom-up analyses

As already discussed in Jokinen & Pelachaud (2013), the *top-down-bottom-up* methodology for data annotation looks at communicative situation from two opposite viewpoints: the top-down approach is based on human observation and uses video recordings and manual tagging of the communicatively important events in the videos, according to some annotation scheme, while the bottom-up approach uses automatic technological means to recognize, cluster, and interpret the signals that the communicating agents emit.

Annotations also need to be consistent, and have the particular semantics they have been designed for, so the annotation results have to be validated by applying the scheme to practical coding tasks and by calculating inter-coder agreement by several coders (including also automatic coding algorithms). By combining the top-down approach, i.e. manual annotation and analysis of the data with the bottom-up analysis of the multimodal signals, it is possible to contribute to the validation of the data and to the quality of the annotated data given the data model and the annotation scheme. On one hand, automatic analysis lends itself to a basis for event detection, and on the other hand, manual annotation is used as a "gold standard" for clustering and classification tasks, to give semantics to the automatically found patterns.

To facilitate manual annotation, semi-manual annotation can be enabled by deploying supervised or unsupervised techniques as a preprocessing step. For instance, speech recognizers can be used to segment speech, parsers to provide linguistic knowledge, eye-trackers to trace gaze paths, and motion trackers as well as various face, gesture and body detection techniques to detect body movement and gesturing. The recognized events can then serve as candidates for more detailed communicative analysis, and the automatic techniques can thus assist human analysis, by segmenting the audio-video data in a uniform manner. Thresholds and parameter values must be set by experimentation and human judgement, but the systematic calculations can be said to produce an objective basis for further analysis, which helps to direct the initial segmentation on same level observations across the annotators, theoretical frameworks, activity types, and conversational settings. Although automatic recognition technologies require a *data model*, i.e. theoretical assumptions that describe the categories and classifications to be found in the data, it is, in principle, easier to determine required granularity and completeness levels by some measurable technical criteria than by more subjective conceptual definitions.

2.2 Internal intention vs external observation

An issue that needs discussion in this context is the very notion of the communicative meaning assumed to be carried by various social signals. It is possible to classify multimodal signals either by interpreting them as originating from the internal communicative intention of the participant, i.e. being displayed or signalled following Allwood's (2001) terminology, or by judging if the events have a noticeable effect on the recipient, i.e. based on the external annotator's observations on what happens in the situation. These two view-points result in different annotations since the former aims to model the participants' internal cognitive decisions, while the latter is based on the results of these actions. A similar distinction can be found in Speech Act theory (Austin 1962), where the notions of illocutionary and perlocutionary acts are introduced. Analogously in multimodal annotation, it is also possible to talk about two different types of annotations, depending on whether the analysis focuses on the *agent's internal intentions*, or on the *consequential effects of the agent's actions* upon the hearer.

2.3 Overt vs Covert meaning

However, even if the annotator is expected to select events that have a communicative function (either by looking at the item's illocutionary or perlocutionary force in the context), there is still another issue that needs attention, namely to determine if the item has an *overt or a covert communicative meaning*. It is well-known that spoken utterances can function as either direct or indirect speech acts, the latter referring to utterances that need deeper contextual inferencing to be correctly interpreted (cf. the classic example of requesting the opening of a window by stating that it is hot), and in a similar manner, multimodal signals can also be regarded as having direct or indirect meanings. For instance, emblems carry a direct, culturally specified meaning (which can be said to be indirect for those outside the culturally specified community), while pointing and iconic gestures directly identify and describe a referent. On the other hand, manipulative gestures, such as

lifting a coffee cup, rolling a pen, changing legs in standing position, etc. do not have an overt communicative function, yet they can indirectly demonstrate the agent's emotional state or intentional stance. They can be appropriately interpreted by the partner only if the partner has learnt to attend to such signalling and is able to draw appropriate conversational inferences to uncover the indirect meanings in the partner's gesturing. To reach the appropriate communicative inferences, the interlocutors need to understand the conversational situation and the principles that guide communication, i.e. they should distinguish the different level of conscious and intentional communication.

2.3 Intentions and segmentation

In human communication, often the difference between unintentional indication (e.g. blushing), intentional display (such as emphasising one's dialect when speaking), and conscious signalling (see Allwood 2001 for terminology) is difficult to determine, since it is difficult to determine the level of consciousness and volition that are behind the communicator's actions in the first place. In general, while it is possible to observe the partner's behaviour and make inferences on the possible reason and motivation for the various actions that the observer considers important in the given communicative situation, it may not be possible to fully understand, nor observe signals and actions in detailed enough manner to actually be able to understand, the actual reasons behind the partner's behaviour.

Human segmentations can thus differ widely depending on also what counts as a relevant event, and behaviour annotations can have different interpretations depending on what aspect of the action the annotator focussed on. The bottom-up approach, or pure signal detection without any particular linguistic knowledge about the meaning of the possible events, may come to help here. While signal analysis can provide rather detailed observations, it can also delay interpretation based on the level of granularity of the data analysis. The relation between form and function need not be one-to-one nor one-to-many, but many-to-many depending on the level of granularity chosen for the analysis in a particular context. The relevance of the various events may become clear only when the data patterns and clusters have been formed, and this can vary depending on the interpretation of the signals.

3 Gesticulation and gesturing

Following Allwood (2001) and Jokinen (2009), we consider interaction as a communication cycle where basic enablements of contact, perception, and understanding must be fulfilled in order for a full communication to take place. Often the enablements are signalled via multimodal signals, which thus form an integral part of the successful communication.

Kendon (2004) uses the term "gesture" to refer to visible action that participants distinguish and treat as governed by openly acknowledged communicative intent. The term "gesticulation" refers to the gesturing that occurs in association with speech and which is bound up with it as part of the total utterance. It consists of three phases (preparatory, peak, and recovery phases) that describe the different parts of the movement.

Interactive gestures form a class with the common function of including the listener in the conversation. They occur at specific moments in time and particular points in space, and can efficiently exert coordination of the conversation and provide meanings as the dialogue goes on.

Gestural signs are formed by the cognitive system that is also used in the movement of the body in the physical environment. Gesturing requires spatio-motoric thinking and ability to orient body parts toward a target in the physical environment, as well as the ability to track the target when it moves (Kita 2000).

Human body movements can be said to form a continuum from movements without any overt communicative meaning to movements which are communicatively significant gesturing. Body movements and the flow of speech are closely linked in human communication system and between the interlocutors. For instance, it is noted that the peak of the gesture coincides with the conceptual focal point of the speech unit, and each new representational gesture appears with a new unit of meaning. Both utterance and gesture derive from a deeper idea unit source that they represent co-expressively.

4 Data and recognition algorithm

For the experimentation we used the 23 dialogues from the MINT (Multimodal INTeraction) project collected at University of Tartu (Jokinen and Tenjes 2012). The speakers are unfamiliar with each other and make acquaintance with their partner for the first time (cf. Paggio et al. 2010).

Each file is about 5 minutes long and records the first encounter between the participants. There are 23 different participant (11 female, 12 male), and each person has dialogues with two different partners, i.e. appears in two videos. The partners face each other, and there are three cameras: one from front and two from sideways recording more on the partner's face from the front. Original Full HD (1920x1080 pixel) videos were resized to 640x360 px and 25 frames per second. A screen shot is given in Figure 1.

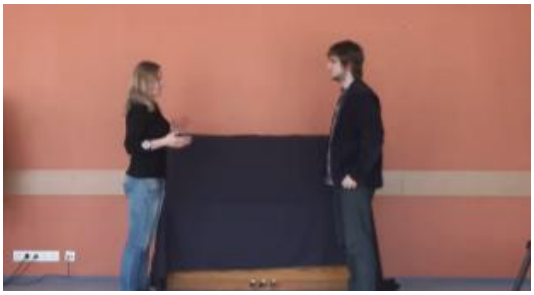

Figure 1 Screen shot from the MINT corpus.

Visual gesture movement recognition consists of several steps (Gonzales and Woods, 2010). On a general level these include:

a) Object representation: compare and decide on the suitable representation for the object tracking. An object can be represented e.g. by its shape or appearance.

b) Feature selection: choose visual features for tracking (colour, texture etc.)

c) Object detection: detect the object based on the chosen features.

d) Object tracking: log the movements of an object by tracking the trajectory of the set of features identified as the object.

e) Object recognition: interpret movements based on the analysis of gathered tracking information.

The previous algorithm (Vels and Jokinen, 2015a) allows us to find the positions of the moving persons in a video frame using a contour detection algorithm. It presented a novel idea of initializing a background model from a single frame using 8-neighbourhood of each of the frame pixel and randomly choosing 20 neighbour-pixels instances to build the model. In the follow-up paper (Vels and Jokinen, 2015b) the contour for the whole body is decomposed into head, torso, and legs bounding boxes so as to allow a more detailed analysis of the movements of the different body parts, by retrieving the precise coordinates of the bounding boxes which can be used to identify hand, head and lower body (foot) movements. Movements are also matched with speech events, which allows correlations to be analysed in easier and improves visualization of the conversational video.

Figure 2 shows a few screen shots of the results of the object segmentation process: background subtraction, morphological closing, body contour, and the final result. The colour video is converted first from RGB to grayscale, the Canny algorithm (Canny, 1986) is used for edge detection, and background subtraction is applied to recognize the objects from the background while morphological closing corrects border areas for final contours.

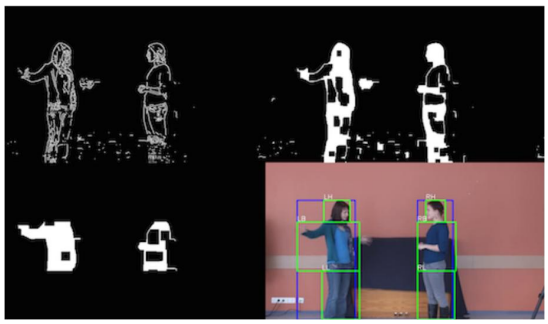

Figure 2 Four of the segmentation steps: background subtraction, closing, body contour, and final result with the detected head, body and leg coordinates.

Decomposition into head, hands, and foot bounding boxes starts by providing a very precise location and size of the head position, and then, using the relative position of the head with respect to the whole body, the body contour is located within which the coordinates for the torso and legs can be retrieved. Hand movement detection uses coordinates with noise removed. Median values of front and back coordinates of body surrounding boxes are used, and all values below a certain threshold are discarded. For the head coordinates, only the middle point value is used, as the head does not change its horizontal size. A simple peak detection algorithm is applied to the coordinates so as to retrieve possible hand movements.

The coordinates are recorded as follows: LBB (left person body back), LBF (left person body front), LH (left person head), RBB (right person body back), RBF (right person body front), and RH (right person head). With these coordinates we can capture all horizontal movements of the human head and body during the conversation.

5 Results and discussion

5.1 Speech and gesturing

The speaker's utterances were annotated by ANVIL manually for three categories: speech, laughs, and non-verbal vocalisations (e.g. *hmm, ahem*). The data is in XML-format and can be parsed automatically for calculating correlation with movement events, and for data visualization.

Figure 3 visualises synchrony of speech with body movements for about a half a minute long clip for the two speakers: the right speaker's movements are shown above, and those of the left speaker below. The right speaker (green coloured above) makes several rapid hand movements (lower green curve) during speaking (light green bars) with two non-verbal vocalisations (topmost dark green bars), but also seems to be rocking his whole body back and forth rhythmically (the green curves move simultaneously and in synchrony with the speech). On the other hand, the left speaker is rather still, and only one significant hand movement (upper blue curve) appears during own speaking (light blue bars). However, the left speaker produces non-verbal vocalisation (dark blue bars) and laughs (top-most dark blue bars) regularly, interleaving them with the partner's speech, and suggesting that the left speaker listens to the partner's lively spoken presentation and gives a lot of feedback to this. This exemplifies cooperation and synchrony between the speakers, and nicely confirms the hypothesis that the speaker moves more than the listener, and that the movements are synchronised (Battersby 2011).

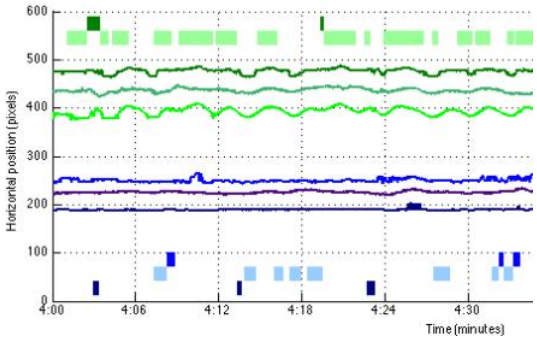

Figure 3 Speech and gesture activities for the left speaker (above) and right speaker (below).

The correlation between speech and gestures is strong, and can be seen in the correlation table, with about 62% of the participants' speech and gestures being in synchrony.

5.2 Movement patterns

Applying the video analysis method to the MINT dataset, we can also get interesting results related to various movement patterns and interaction synchrony among the conversation partners. As shown in Vels and Jokinen (2015b), a variety of gestures can be recognized by their combined movement curves, i.e. it is possible to recognize certain type of gesturing based on their characteristic bounding box trajectories. For instance, Figure 4 exemplifies beat gesturing, i.e. rhythmic hand gesturing during one's own speech, and clearly shows the variation in the front coordinates of the bounding box corresponding to hand movements. Figure 5 shows how a whole body moving forward provides a simultaneous set of back- and forward pikes in the curves related to upper body (hands) and lower body. Figure 6 shows how a large spike in the back coordinate of the body bounding box without movement in the head or the front coordinates of the body bounding box imply that there is gesturing behind the speaker. Finally, Figure 7 shows that if there are spikes both in the front and back coordinates of the body bounding box but the head coordinates is unchanged, the speaker waves her hands around.

6 Conclusion

We have discussed automatic recognition of human body movement and its use in communication studies. We used our previous algorithm for detecting body movement on video films and especially the version that can distinguish the three parts of the body: the head, the torso and the legs. We applied top-down-bottom-up approach and confirm earlier hypothesis of the gesturing and body movement as activities closely related to speaking.

The results show that the method can be applied with fairly good results, and combining the movement with speech occurrences we can visualise the interaction and especially the synchrony with speech and gestures. This is analogous to Campbell and Scherer (2010) who measured synchrony and alignment in spoken interactions, or Jokinen (2009) who applied the same method to measure conversational activity.

Future work concerns more detailed analysis of the MINT dataset and improving the hand and head movement detection algorithm. We will also use the same algorithm on other corpora and compare the gesturing in the context of intercultural communication. From the detected

body movements it is also interesting to try to extract gestures and their interpretation automatically. It is expected that the research presented in this paper can be used to integrate the user's body movement with the autonomous agent's gesture recognition capability, so as to produce natural interaction, and the models built using the help of bounded boxes and their visualisation as graphs will help to design the agent's own gesture model to produce appropriate gesturing and gesticulation in the course of the interaction.

Acknowledgement

The work has been done within the Estonian Science Foundation project MINT *Multimodal Interaction* (ETF 8958) and the project IUT 20-56 *Computational Models of Estonian*. I wish to thank Heiki-Jaan Kaalep for support, and Martin Vels for implementing the recognition algorithm.

References

Allwood, J. 2001. Dialog Coding—Function and Grammar. Gothenburg Papers. Theoretical Linguistics, 85. Department of Linguistics, Gothenburg University.

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta and P. Paggio. 2007. The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In Martin, J.C. et al. (eds.), Multimodal Corpora for Modelling Human Multimodal Behaviour. Special issue of the International Journal of Language.

Argyle, M. 1988. Bodily Communication. London: Methuen.

Argyle, M., Cook, M. 1976. Gaze and Mutual Gaze. Cambridge University Press.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford University Press.

Battersby, S. 2011. Moving Together: the organization of Non-verbal cues during multiparty conversation. PhD Thesis, Queen Mary, University of London.

Campbell, N., Scherer, S. 2010. Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity. *Proceedings of Interspeech*. Makuhari, Japan

Canny, J. 1986. A Computational Approach to Edge Detection, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6), pp. 679-698.

Endrass, B., Rehm M., Andre, E. 2009. Culture-specific communication management for virtual agents. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09). 281–288

Gonzales, R. C., Woods, R. E. 2010. *Digital Image Processing* (3rd edition). Pearson Education, pp. 652-661.

Jokinen, K. 2009. Gestures in Alignment and Conversation Activity. *Proceedings of the Conference of the Pacific Association for Computational Linguistics Conference (PACLING),* Sapporo, Japan, pp. 141-146.

Jokinen, K. 2009. *Constructive Dialogue Modelling:Rational Agents and Speech Interfaces*. Chichester: John Wiley.

Jokinen, K., Pelachaud, C. 2013. From Annotation to Multimodal Behavior. In Rojc, M. & Campbell, N. (Eds.) Co-verbal Synchrony in Human-Machine Interaction. CRC Press, Taylor & Francis Group, New York.

Jokinen, K., Tenjes, S. 2012. Investigating Engagement – Intercultural and Technological Aspects of the Collection, Analysis, and Use of Estonian Multiparty Conversational Video Data. *Proceedings of LREC'12*, pp. 2764 – 2769. Istanbul, Turkey: ELRA.

Kendon, A. 2004. *Gesture: Visible action as utterance.* New York: Cambridge University Press.

Kita, S. 2000. How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture*. pp. 162-185. Cambridge: Cambridge University Press. Open Access accepted version: http://wrap.warwick.ac.uk/66257/

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., Navarretta, C. 2010. The NOMCO Multimodal Nordic Resource – Goals and Characteristics. In *Proceedings of LREC'10,* Valetta, Malta: ELRA.

Schuller, B, Batliner, A. 2013. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley, 2013.

Vels, M., Jokinen, K. 2015a. Recognition of Human Body Movements for Studying Engagement in Conversational Video Files. In: Jokinen, K. & Vels, M. (eds) Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014. Tartu, Estonia. 110:014. Linkping: LiU Eletronic Press.

Vels, M., Jokinen, K. 2015b. Detecting Body, Head, and Speech in Engagement. *Proceedings of the IVA 2015 Workshop on Engagement in Social Intelligent Virtual Agents* (*ESIVA 2015).*

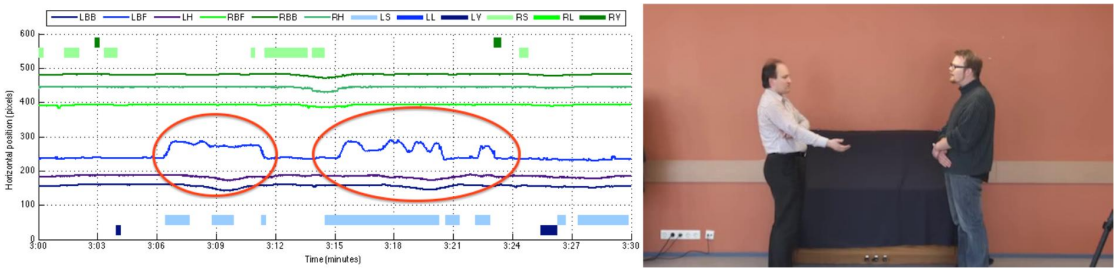

Figure 4 Beat gesturing simultaneous with speech (indicated by a light blue bar underneath the movement).

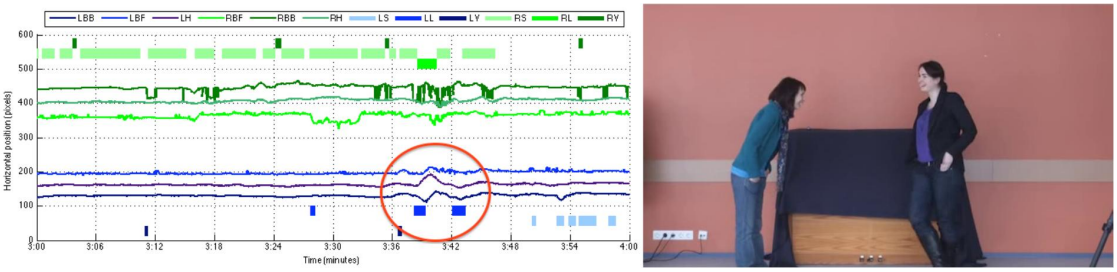

Figure 5 Leaning body movements in between speech and simultaneous with the partner's speech and gesturing.

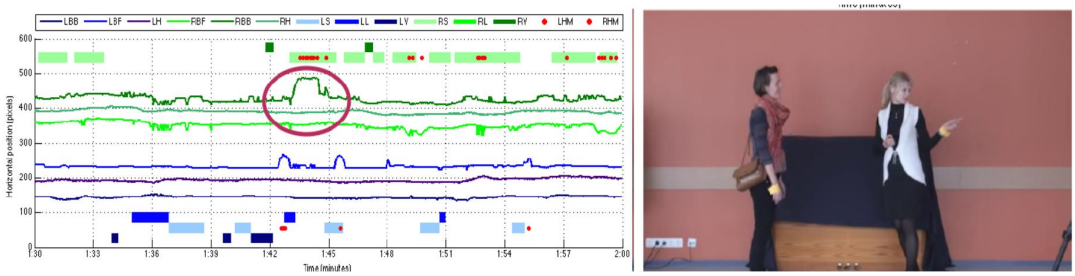

Figure 6 Large spike in the RBB coordinate without the RH or RBF => gesture somewhere behind the speaker.

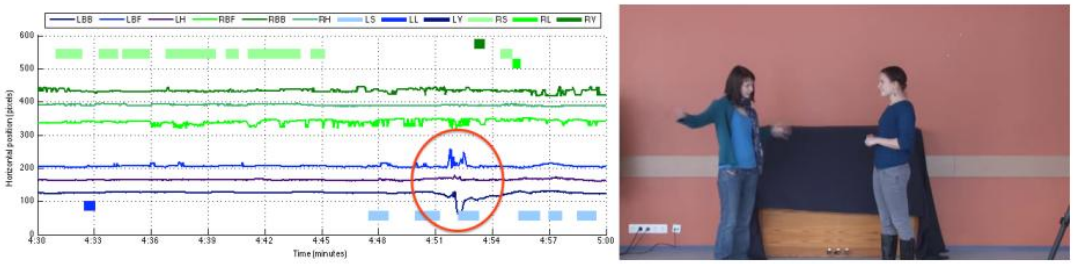

Figure 7 Spikes in both LBR and LBF coordinates with unchanged LH coordinate => the speaker waves her hands around