MOTION CAPTURE IN GESTURE AND SIGN LANGUAGE RESEARCH

Mehmet Aydın Baytaş
Qualisys AB, Koç University
mbaytas@ku.edu.tr

Damla Çay
Koç University
dcay13@ku.edu.tr

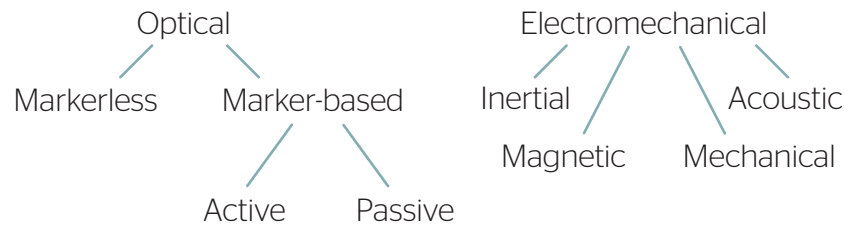
Asım Evren Yantaç
Koç University
eyantac@ku.edu.tr

Morten Fjeld
Chalmers University of Technology
fjeld@chalmers.se

Motion capture enables **precise**, **quantitative** analysis of gesture and sign language production.

To introduce a wider audience of researchers to this field of inquiry, we present a review previous works that utilized motion capture to study sign and gesture production, along with comments on technical and methodological issues.

There exist a wide variety of methods and hardware for motion capture.



SIGN LANGUAGES

While earlier motion capture studies on sign languages of the deaf have focused on single signs or short sequences (Wilbur, 1990; Wilcox, 1992), more recent work deals with continuous, longer-duration, conversational data, and aims at supporting automated transcription and translation, as well as synthesis.

ARTICULATION AND PROSODY

Hypoarticulation in American Sign Language (ASL) production has been studied extensively through motion capture. Signing speed and the locations of adjacent signs (**Mauk et al., 2008**), as well as a sign's position within an utterance (**Tyrone & Mauk, 2010**) have been found to influence the clarity of articulation. Further studies have explored directionality in coarticulation effects, effects of body posture, and the use of signing space (**Mauk & Tyrone, 2012**) in relation to hypoarticulation. **Tyrone et al. (2010)** studied variations in sign prosody induced by the location of a particular sign in a phrase, finding similarities between signed and spoken languages, and support their hypothesis that a similar framework can apply to both. **Puupponen et al. (2015)** investigated the functions and kinematics of head movements in Finnish Sign Language (FinSL), but found "non-categorical" relationships between form and function. **Tyrone and Mauk (2016)** reported on the phonetic role of non-manual articulators (head and body movements).

SYNTHESIS

Synthesis (or generation) pertains to sign production by animated (3D, cartoon, or robotic) characters.. **Lu & Huenerfauth (2010)** describe the design of an ASL database to support synthesis, as well as an evaluation of their design and recording methods. In later work, they use data from an ultrasonic/inertial hybrid mocap system to train a vector-based language model to improve the understandability of inflecting verbs. The *SignCom* project (**Duarte & Gibet, 2010a/2010b; Gibet et al., 2011**) aims to support generation by contributing phonetic analyses of French Sign Language (LSF) based on mocap and video data. Using methods from linguistics and computer animation, *SignCom* researchers adopt a "target-based" view of sign production where signs are considered to be "sequences of targets" and signers "improvise" transitions between targets. The more recent *Sign3D* project (**Lefebvre-Albaret et al., 2013**) aims to address related challenges in database design, data retrieval, rendering, and user-friendly tools to improve the workflow.

TECHNICALITIES

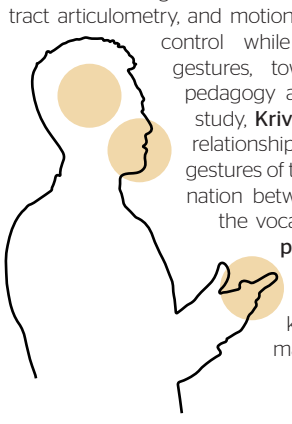
When recording sign languages with marker-based optical mocap, there is a trade-off between the level of detail to be captured, and the comfort of the subject. Too many markers can interfere with articulation, and too few markers can require extensive post-processing or simply do not capture the required information. **Jantunen et al. (2012)** report on their experiences in collecting and processing marker-based motion capture for FinSL research. They offer strategies for working with trade-offs in system setup, marker placement, and post-processing automation. Technical issues are also important for electromechanical motion capture, such as the challenge of properly fitting these sensors to the hands of deaf subjects (**Lu & Huenerfauth, 2009; Huenerfauth & Lu, 2010**). **Tyrone (2015)** discusses these issues and other considerations for instrumented studies of sign production. More recently, **Jantunen et al. (2016)** describe the creation of an annotated FinSL corpus using Kinect and computer vision data, leveraging recent developments in markerless mocap.

CO-SPEECH GESTURES

Motion capture studies on co-speech gestures often aim to understanding gestures in everyday communication, language development, and speech/language impairments. More recent studies have also been motivated by topics in computing: activity and affect recognition, machine translation, natural avatar animation, and improving multimodal data analysis methods.

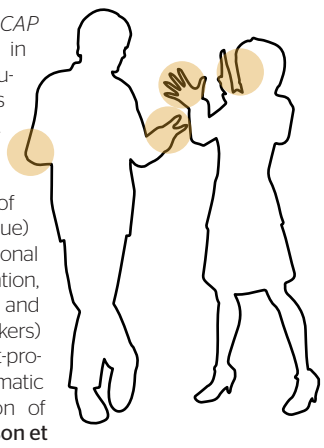
PROSODY AND MULTIMODAL DATA

Krivokapic, Tiede & Tyrone (2015) have investigated the relationship between prosodic structure and large-scale body movements.Using multimodal data (audio, electromagnetic vocal tract articulometry, and motion), they analyze whether prosodic control while speaking extends to bodily gestures, towards understanding language pedagogy and speech pathology. In a later study, **Krivokapic et al. (2016)** explored the relationships between deictic gestures and gestures of the vocal tract, and found a coordination between the "intonation gestures" in the vocal tract and the deictics. **Krivokapic, Tiede & Tyrone (2017)** presented further studies of the effects of prosodic structure on the kinematics of both speech and manual gestures.



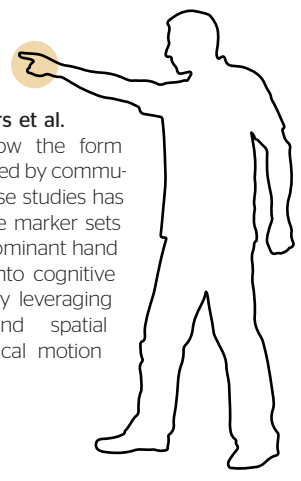
DIALOGUE

Busso et al.'s (2008) *IEMOCAP* dataset of actors in emotion-based dyadic communication scenarios supports HCI and linguistics research. **Edlund et al.'s (2010)** *Spontal* dataset (60+ hours of audio, video, and mocap recordings of spontaneous Swedish dialogue) is for studying conversational phenomena (e.g. floor negotiation, turn taking, feedback, and synchrony between speakers) (**Beskow et al., 2011**) and post-processing methods (e.g. automatic segmentation and annotation of multimodal data) (**Alexanderson et al., 2013**). **Alexanderson et al. (2014)**, for example, used machine learning to infer, from *Spontal* motion data, whether or not a person is speaking; while also revealing the saliency of various features for the task.



INTENTION AND DEICTICS

Bonfiglioli et al. (2009) have used motion capture to study the conceptualization of deictic pronouns and their effects on movement planning. **Sartori et al. (2009)** have demonstrate the effect of communicative intent on the kinematics of functional gestures, while **Peeters et al. (2013)** similarly investigated how the form pointing gestures can be influenced by communicative intent. Data for all of these studies has been collected using very simple marker sets comprising 1-3 markers on the dominant hand only. Using this data, inquiries into cognitive processes have been possible by leveraging the superlative temporal and spatial resolution of marker-based optical motion capture.



TECHNICAL CONSIDERATIONS AND LIMITATIONS

COST

Marker-based and electromechanical motion capture systems that afford high spatial and temporal resolution can be expensive. Cheaper systems such as depth cameras that detect human joint positions may not provide the resolution or accuracy that researchers require. Achieving high-resolution motion capture with low-cost hardware is an **active field of research** (e.g. Elhayek et al., 2017).

NOISE

Optical mocap systems can be susceptible to interference from infrared light (e.g. sunlight) and reflective materials in the scene. Other systems may be sensitive to various sources of noise such as electromagnetic emissions, light, or sound. **Environmental factors** must be considered when designing motion capture studies.

SKILLS

Collecting data with almost any motion capture system requires researchers to be knowledgeable about its technicalities, as well as the problem domain. The physical capabilities and limitations of the system are important considerations, along with the annotation, indexing, retrieval, and analysis of data. Conducting sign and gesture research with motion capture **requires interdisciplinary learning and/or collaboration**.

ACCURACY AND PRECISION

While marker-based optical systems can make position measurements with sub-milimeter accuracy and precision, joint positioning errors in state-of-the-art markerless motion capture of human subjects are currently on the order of centimeters (Elhayek et al., 2017). Electromechanical systems differ broadly in terms of their accuracy measures. Care must be taken to **ensure that the measurement accuracy and precision are appropriate** for the inquiry at hand,

OCCLUSIONS

Optical motion capture systems require a **clear line of sight** between the camera(s) and the markers or bodies of interest. Anything that occludes the capture subject, even partially, may interfere with the measurement. Increasing the number of cameras may mitigate occlusion issues. For this reason, while researchers studying pointing and grasping with minimal marker sets can work with as few as 3 cameras, it is common to see 8- or 12-camera setups for recording sign languages or natural conversation.

SAMPLING RATE AND RESOLUTION

A high sampling rate and resolution can capture even fast and minute movements in detail. The trade-off is that the size of the data will increase with the sampling rate, and only costly high-end mocap sytems will be able to provide such temporal resolution. It's also possible that different software in the analysis pipeline can impose limits on file size and throughput. The sampling rate and resolution for the capture must be **selected in an informed fashion and confirmed with pilot studies**.

REFERENCES

Alexanderson, S., House, D., & Beskow, J. (2013). Extracting and analysing co-speech head gestures from motion-capture data. *Proc. Fonetik 2013*.
Alexanderson, S., Beskow, J., & House, D. (2014). Automatic speech/non-speech classification using gestures in dialogue. *Proc. Swedish Language Technology Conference*.
Beskow, J., Alexandersson, S., Al Moubayed, S., Edlund, J., & House, D. (2011). Kinetic data for large-scale analysis and modeling of face-to-face conversation. **Proc. AVSP**.
Bonfiglioli, C., Finocchiaro, C., Gesierich, B., Rositani, F., & Vescovi, M. (2009). A kinematic approach to the conceptual representations of this and that. *Cognition*, 111(2), 270-274.
Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335.
Duarte, K., & Gibet, S. (2010a). Heterogeneous Data Sources for Signed Language Analysis and Synthesis: The SignCom Project. *Proc. LREC 2010*.
Duarte, K., & Gibet, S. (2010b). Reading between the signs: How are transitions built in signed languages. *Proc. TISLR 2010*.
Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture. *Proc. LREC 2010*.
Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., ... & Theobalt, C. (2017). MARCOnI—ConvNet-Based MARker-Less Motion Capture in Outdoor and Indoor Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(3), 501-514.
Gibet, S., Courty, N., Duarte, K., & Naour, T. L. (2011). The SignCom system for data-driven animation of interactive virtual signers: methodology and evaluation. *ACM TiiS*, 1(1), 6.
Huenerfauth, M., & Lu, P. (2010). Accurate and accessible motion-capture glove calibration for sign language data collection. *ACM TACCESS*, 3(1), 2.
Jantunen, T., Burger, B., De Weerdt, D., Seilola, I., & Wainio, T. (2012). Experiences from collecting motion capture data on continuous signing. *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon* (pp. 75-82).
Jantunen, T., Pippuri, O., Wainio, T., Puupponen, A., & Laaksonen, J. (2016). Annotated Video Corpus of FinSL with Kinect and Computer-Vision Data. *Proc. 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proc. LREC 2016*.
Krivokapic, J., Tiede, M., & Tyrone, M. E. (2015). Kinematic properties of concurrently recorded speech and body gestures and their relationship to prosodic structure. *The Journal of the Acoustical Society of America*, 137(4), 2269-2269.
Krivokapic, J., Tiede, M., Tyrone, M. E., & Goldenberg, D. (2016). Speech and manual gesture coordination in a pointing task. *Proc. Speech Prosody 2016*.
Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection. *Laboratory Phonology*, 8(1).
Lefebvre-Albaret, F., Gibet, S., Turki, A., Hamon, L., & Brun, R. (2013). Overview of the Sign3D Project High-fidelity 3D recording, indexing and editing of French Sign Language content. *Proc. SLTAT 2013*.
Lu, P., & Huenerfauth, M. (2009). Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects. *Proc. SIGACCESS 2009* (pp. 83-90).
Lu, P., & Huenerfauth, M. (2010). Collecting a motion-capture corpus of American Sign Language for data-driven generation research. *Proc. NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies* (pp. 89-97).
Lu, P., & Huenerfauth, M. (2012). Learning a vector-based model of American Sign Language inflecting verbs from motion-capture data. Proc. *Third Workshop on Speech and Language Processing for Assistive Technologies* (pp. 66-74).
Mauk, C. E., & Tyrone, M. E. (2012). Location in ASL: Insights from phonetic variation. *Sign Language & Linguistics*, 15(1), 128-146.
Mauk, C. E., Tyrone, M. E., Sock, R., Fuchs, S., & Laprie, Y. (2008). Sign lowering as phonetic reduction in American Sign Language. *Proc. 2008 International Seminar on Speech Production* (pp. 185-188).
Peeters, D., Chu, M., Holler, J., Ozyurek, A., & Hagoort, P. (2013). Getting to the point: The influence of communicative intent on the kinematics of pointing gestures. **Proc. CogSci 2013** (pp. 1127-1132).
Puupponen, A., Wainio, T., Burger, B., & Jantunen, T. (2015). Head movements in Finnish Sign Language on the basis of Motion Capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls. *Sign Language & Linguistics*, 18(1), 41-89.
Sartori, L., Becchio, C., Bara, B. G., & Castiello, U. (2009). Does the intention to communicate affect action kinematics?. *Consciousness and Cognition*, 18(3), 766-772.
Tyrone, M. E. (2015). Instrumented Measures of Sign Production and Perception. In *Research Methods in Sign Language Studies: A Practical Guide*, 89-104.
Tyrone, M. E., & Mauk, C. E. (2010). Sign lowering and phonetic reduction in American Sign Language. *Journal of Phonetics*, 38(2), 317-328.
Tyrone, M. E., & Mauk, C. E. (2016). The Phonetics of Head and Body Movement in the Realization of American Sign Language Signs. *Phonetica*, 73(2), 120-140.
Tyrone, M. E., Nam, H., Saltzman, E., Mathur, G., & Goldstein, L. (2010). Prosody and movement in American Sign Language: A task-dynamics approach. *Proc. Speech Prosody 2010*.
Wilbur, R. B. (1990). An experimental investigation of stressed sign production. *International Journal of Sign Linguistics*, 1(1), 41-60.
Wilcox, S. (1992). *The phonetics of fingerspelling*. John Benjamins Publishing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676063