Title: Deciphering the Communicative Code in Speech and Gesture Dialogues by Autoencoding Hypernetworks

Authors:
N.-S. Nam[1,2], K. Bergmann[1], U. Waltinger[1], S. Kopp[1], I. Wachsmuth[1], B.-T. Zhang[1,2]
SFB 673 and CITEC, Bielefeld University[1]
CBIT and ICS, Seoul National University[2]

Text:

What kinds of grammar or code are used in interactive communications with speech and gestures? How varied or invariant is this code among people in a language community? What types of communicative code facilitate the alignment of the speech and gesture for language understanding?

To study these and other related questions we develop computational techniques using coding theory and machine learning that decipher the communicative code in embodied multimodal interaction. We use data from the SaGA (Bielefeld Speech and Gesture Alignment) corpus which consists of 25 dyads of naturalistic, yet controlled, and systematically annotated speech and gesture use, engaged in a spatial communication task (Luecking, 2010). For the work we present here, a sub-corpus of 5 dyads is employed (473 noun phrases, 288 gestures) combining three kinds of information. First, gesture coding including gestural representation techniques (e.g., drawing, placing) and morphological gesture features (e.g., handshape). Second, a transcription of the spoken words and dialogue contextual information (information state, thematization, elemental actions of direction giving). And third, a coding of the gestures' referent objects and their spatio-geometrical properties (dimensionality, symmetries, etc.).

We formulate the gesture generation problem as an encoding problem and use the unsupervised, autoencoding technique, where the input vector $\mathbf{x}$ is transformed by some function $f(.;\mathbf{W})$ to the output vector $\mathbf{y}$ which is the same as the input, i.e. $\mathbf{y} = f(\mathbf{x};\mathbf{W}) = \mathbf{x}$. For transformation we use the hypernetwork graphical architecture. The hypernetwork is a hypergraph structure, where the edges are weighted and represent the subsets of the variables (variables). One advantage of the hypernetwork is that it can capture the compositional structures or *code words* (or construction grammar rules) in its hypergraph structure. We apply an expectation-maximization style of learning algorithm to build the best autoencoding hypernetwork for the observed gesture-speech dialogue data. Another advantage of the hypernetwork is its generativity, i.e. the hypernetwork model can generate the values of the unknown (unobserved) variables from those of the known (observed) variables by probabilistic inference. This feature is especially useful for artificial communicative agents since the learned hypernetwork can be used to synthesize the gestures for virtual avatars or humanoid robots.

Acknowledgments:

This work was supported by the National Research Foundation (NRF) grants (2011-0016483-Videome, 2010-0018950-BrainNet), the IT R&D Program of KEIT (10035348-mLife), and the BK21-IT Program.

References:

Bergmann, K., & Kopp, S. (2010). Modelling the production of co-verbal iconic gestures by learning Bayesian decision networks. *Applied Artificial Intelligence* 24(6):530–551.
Kopp, S., Bergmann, K. & Wachsmuth, I. (2008). Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production. *Semantic Computing* 2(1):115–136.
Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In *LREC 2010 Workshop.*
Zhang, B.-T. (2008). Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory. *IEEE Comp. Intell. Mag.*, 3(3), 49-63.
Zwann, R. A. & Kaschak, M. P. (2008). Language in the brain, body, and world. Chap. 19, *Cambridge Handbook of Situated Cognition.*

Keywords: alignment of speech and gesture, communicative codes, construction grammars, autoencoding hypernetworks, unsupervised learning.

Wordcount: 499