Metaphoric gestures: towards grounded mental spaces

Margot Lhommet and Stacy Marsella
Northeastern University
360 Huntington Avenue - Boston, MA, USA
[lhommet|marsella]@neu.edu

Abstract. Gestures are related to the mental states and unfolding processes of thought, reasoning and verbal language production. This is especially apparent in the case of metaphors and metaphoric gestures. For example, talking about the importance of an idea by calling it a big idea and gesturing to indicate that large size is a manifestation of the use of metaphors in language and gesture. We propose a computational model of the influence of conceptual metaphors on gestures that maps from mental state representations of ideas to their expression in concrete, physical metaphoric gestures. This model relies on conceptual primary metaphors to map the abstract elements of the mental space to concrete physical elements that can be conveyed with gestures.

Keywords: Nonverbal behavior, gesture, metaphor, embodied cognition, embodied conversational agent

1 Introduction

Gestures play a powerful and diverse role in face-to-face interaction. They are meaningfully related to the structure of mental states and unfolding processes of thought.

Our work focuses on a generative model of gesturing that allows virtual humans to communicate by using multimodal behaviors including speech, gesture and other nonverbal behaviors such as gaze, posture shifts or facial expressions.

When studying the relation between mental state and gestures, a key challenge arises: gestures' form and meaning are largely improvised and understood in context. *Emblems* have highly conventionalized meaning (such as the "thumb-up" gesture that means "okay" [9]). *Iconic* gestures detail the mental image conveyed by the speaker by depicting properties or actions taken on objects, such as their size or mimicking their movement. *Deictics* consist in pointing at world locations and are particularly useful to disambiguate references to objects and locations. Gestures can also be physical manifestations of abstract concepts, showing the size of ideas to represent their importance or locating events on a time line as if they were objects in space. Once materialized, physical actions can be taken on these objects such as rejecting an idea by a sideways flip of the hand [2]. These gestures are called *metaphoric* since they consider abstract objects "as-if" they were concrete objects.

Gestures' meaning is manifold and context dependent. In their realization, however, gestures are physical actions in the speaker's immediate physical environment, inherently described in physical terms such as size, location or path.

Our long-term research aims to generate verbal and nonverbal behavior that realize specific communicative intentions from a speaker's mental space. In this paper, we focus on the generation of metaphorical gestures[1]. How can physically constrained gestures express such a wide range of mental states?

Indeed, speech and gesture don't reflect the entirety of ongoing thoughts. A speaker's discourse follows a flow of ideas, combining speech and nonverbal behaviors to convey certain intentions and facilitate a listener's understanding. What is actually expressed is a specific part of the mental state, a *mental space*, a "partial and temporary structure which speakers construct when thinking or talking about a perceived, imagined, past, present or future situation."[8, p.3][2]. How this mental space is built and what it contains depends on the context and on the communicative intentions of the speaker.

Our model draws inspiration from embodied cognition that suggests that we use the same set of sensory and motor representation to make sense of our world. Cognitive linguists proposed the Conceptual Metaphor Theory according to which we understand abstract concepts by mapping them to concrete elements by using *conceptual metaphors* (sometimes called *image schemas*) [12]. For example, we make sense of a "big idea" by mapping the importance of an idea (an abstract property of an abstract object) to the size of a concrete object.

We propose that, to be expressed with gestures, the mental space has first to be conceptually grounded, i.e. mapped to concrete elements from which they inherit physical properties. These properties are then combined into gestures that convey the desired communicative intentions. Such a model supports a generative model of gesturing that:

1. allows for a large space of mental representations to be mapped to a comparatively small space of metaphoric gestures,
2. can convey complex communicative intentions via composition over this small set of gestures,
3. guides how properties in abstract propositions (such as "important idea") can be conveyed by manipulations of the gestures (big gesture).

After describing the components of our model, we detail its current implementation and illustrate it with examples. Finally, we comment on the implementation and discuss future work.

2 Related work

Researchers have explored several techniques to automate the generation of virtual humans' nonverbal behaviors that realize communicative intentions.

Most approaches take speech as input to generate appropriate nonverbal behavior, but they differ on how the models were developed, the degree of automation in the generation process itself and the particular classes of nonverbal behaviors that are handled. Specifically, some systems use annotated text that specifies what information has to be conveyed nonverbally (e.g. [11]). Such ap-

[1] See the discussion for an account of the generation of multimodal behavior.

[2] A mental space is similar to McNeill's Growth Point, "a minimal unit of dialectic in which imagery and linguistic content are combined."[18, p.18]

proach is cumbersome since it requires manual annotations of the utterance's text. Data-driven techniques can automate the generation of specific classes of nonverbal behaviors from specific input. For example, prosody has been used to generate gestures [16] and text has been mapped to head movements [14] and gesturing style [10, 19]. Another approach consists in analyzing the speech to infer the underlying communicative intentions. BEAT infers rheme and theme from the text to generate intonation and emphasis [3]. NVBG detects communicative functions in the text (e.g. affirmation, emphasis, disfluencies) based on a keywords mapping [15]. Cerebella integrates acoustic, syntactic and semantic analyses to infer communicative intentions and elements of the mental state (emotional state, energy, emphasis,. . .) [17]. The common critique is that while deeper and more elaborate analyses allow inferring and conveying the communicative intentions present in the speech, the nonverbal behavior generated is limited in the range of what can be inferred from the speech utterance only.

This can be overcome by integrating deeper cognitive processes that co-generate speech and gesture. [1] study the co-production and coordination of speech and gesture production under linguistic and cognitive constraints. In particular they show how the conceptualization of path, motion and manner constrain speech and iconic gesture production. [13] formalize the relation of gesture and speech with a logical form of multimodal discourse, in particular between a discourse's spatial elements and deictic gestures.

Our goal is the co-generation of speech and gesture based on a common representation of the communicative intentions. Therefore, our work investigates the content of this underlying common representation (the mental space) and the processes that map it to speech, gestures and nonverbal behaviors. In particular, we propose to explicitly represent the mental space and its grounded counterparts, that allows to combine its expression through multiple channels as well as representing sequences of actions taken on existing elements. In this paper, we investigate the generation of metaphorical gestures.

3 Model

The Figure 1 presents the elements involved in our model. To generate a gesture plan according to a mental space, we propose to ground this mental space in concrete domains by using primary metaphors. Primary metaphors are conventional mappings that associate elements from abstract domains to elements in the concrete domain [12, 8]. "There are hundreds of such primary conceptual metaphors, most of them learned unconsciously and automatically in childhood simply by functioning in the everyday world with a human body and brain. There are primary metaphors for time, causation, events, morality, emotions, and other domains that are central to human thought" [12, p.257].

Mental Space What the processes of thought and reasoning look like, what the content of the mental space is, are open questions that we do not claim to answer in their entirety. We define a mental space as a structure that reflects communicative intentions: it contains the information to express, as well as information regarding how to express it or modify something previously said.

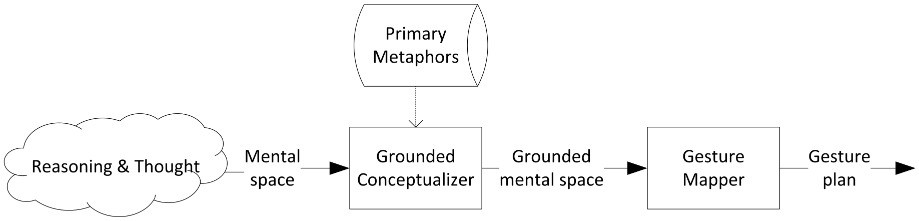

Fig. 1: Elements and processes of the model.

Grounded Conceptualizer The Grounded Conceptualizer maps the abstract elements of a mental space to concrete physical elements. This mapping consists in a systematic projection of the objects, properties and relations from one domain to another and is based on primary metaphors. Once grounded, the abstract elements of the mental space have concrete counterparts that can be expressed with gestures and the communicative intention is mapped into actions on these concrete objects.

Gesture Mapper Gestures are physical actions on the speaker's immediate physical environment. Therefore, gestures are inherently described in physical terms such as size, location, path... The Gesture Mapper generates a gesture plan by combining the concrete elements of the grounded mental space according to the communicative intention.

4 Implementation

In this paper, we formalize the relation between an abstract idea and a concrete object, and a set and a container, respectively following the metaphors Abstract Object is Concrete Object and Set is a Container). These two metaphors seem to be the root elements involved in many primary metaphors [12, 8]. Indeed, the concrete domains used in metaphorical mappings can be more elaborated but in this paper we explore the range of metaphors and metaphorical gestures that we can reach with these two.

In this section, we detail the current implementation of our model. We use an example that comes from a role-play conversation between a clinician (the speaker) and an actress pretending to suffer from PTSD (Figures 2-6).

4.1 Mental Space

The two main concepts used in this paper are *elements* and *sets*. Elements represent abstract concepts. they have a type, some properties and can be associated to other elements via predicates. Sets are structures that group elements sharing a similar property. Mental spaces are described using OpenCyc[3].

In this paper, we demonstrate the implementation of three communicative intentions:

- Depict a property of an element: *depict_property(e p)*

[3] OpenCyc is a large ontology-like knowledge base. Collections are hierarchically organized and properties can be inherited. `http://www.cyc.com`

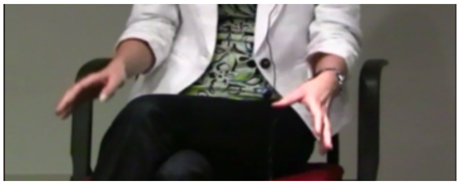

Fig. 2: Depict the cardinality of a set

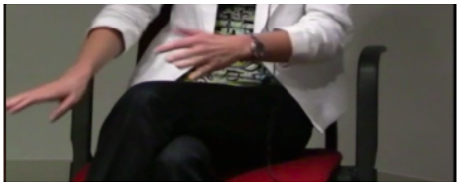

Fig. 3: Remove an element from a set

- Manipulate an existing set[4] to add or remove elements: *add(e s) remove(e s)*.
- Contrast the difference of value between a property of two elements[5]: *contrast(p e1 e2)*

As stated above, the content of the mental space is driven by what information is required to convey a specific intention with metaphoric gestures. The information required for these communicative intentions is given below.

Depict a property

```
(operation depict_property(s p))
(isa s set)
(p s value)
```

Remove an element

```
(operation remove(s e))
(isa e element)
(isa s set)
(contains s e)
```

Contrast two elements

```
(operation contrast(p e1 e2))
(isa p property)
(isa e1 element)
(isa e2 element)
(f e1 value1)
(f e2 value2)
```

Example The clinician asks: “(1) Is there anything, (2) besides what he wants, anything you want to work on?”On the first part of the sentence (1), she offers to discuss any thing and depicts the cardinality of a large set that contains all the possible discussion topics (see Figure 2). On the second part of the sentence (2), she realizes that there is a topic included in this set that she does not want to discuss (“what your husband wants”) so she removes it from the set, generating a similar action on the container created earlier (see Figure 3). These mental spaces are described below.

Depict cardinality

```
(operation depict_property(s cardinality))
(isa s set)
(cardinality s count(s contains t))
(forall t
  (and (isa t element)
       (type t discussionTopics)
       (contains s t)))
```

Remove an element

```
(operation remove(s e))
(isa e element)
(isa s set)
(contains s e)
(type e discussionTopics)
(wants husband e)
```

[4] The grounded elements of the mental space are persistent over time.

[5] For clarity we only show a contrast between two elements, but the notation could be extended to consider more of them.

4.2 Grounded Conceptualizer

The Grounded Conceptualizer maps the abstract elements of the mental space to physical properties that convey the same meaning. As mentioned above, we particularly study the domains of *Object* and *Container*. An *Object* has physical properties such as a size, a weight, a location . . . A *Container* is an object that contains other objects.

A set of rules, based on primary metaphors, maps the content of the mental space to physical objects and properties. The following rules state that an element is mapped to an object and a set to a container.

```
(type e element) --> (type e’ object)
(type e set) --> (type e’ container)
```

OpenCyc uses a large hierarchy of collections to represent concrete elements such as Location, Path, Shape. . . that we use to represent the physical properties of objects and containers. Once the mapping space is grounded, the physical knowledge is leveraged by rules that represent the logics of the physical world. These objects are created in a specific micro-theory inside OpenCyc to control the inferences that are made. For example, the rule below associates a container’s location to the objects it contains.

```
(type c container) (location c loc) (contains c obj) --> (location obj loc)
```

Example To illustrate this grounding process, we show how the first part of the sentence cited earlier is grounded and what information is elaborated. The set of possible topics is mapped to a container that contains all topics objects. The operation consisting in depicting the cardinality of the set is mapped to an operation depicting the size of the bound container. A physical inference is made: the size of a container depends on what it contains; here, the whole set of topic objects, and is therefore big.

```
(type s’ container)                  | (operation depict_property(s’ size))
(contains s’ (all (type  t’ object)))| (size s’ big)
```

4.3 Gesture Mapper

Generic gesture rules combine the physical properties of the grounded mental space according to the desired communicative function. They are explained below and illustrated on the previous examples.

Depict a property Because they are grounded, all abstract properties (such as the importance or the cardinality of a set) are mapped to physical elements (for example, size, location, shape, weight,. . .). The physical properties of objects are retrieved to generate a gesture. In particular, the salient property to express (p) and its value *val* are added to the gesture specification. Other physical properties are also added to further refine the gesture, but they are optional. For example, the first part of the clinician sentence (i.e. “Is there anything you want to talk about ?”) is transformed into the following structure:

```
<goal=depict id=Container01 type=container size=big/>
```

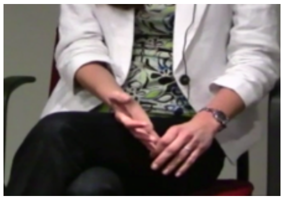

Fig. 4: Enumeration

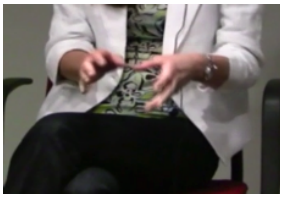

Fig. 5: Contrast now

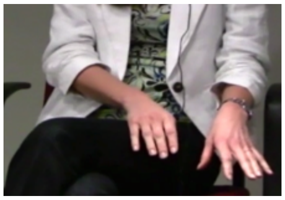

Fig. 6: Contrast past

Add or remove an element from a set This communicative intention implies that the set (and the bound container) already exist in the mental space (as well as in the physical gesture space, i.e. they have a location). The gesture rule to remove an element from a container is:

```
(operation remove(s o)), (isa s container), (location s loc)
--> <goal=remove source_location=loc target=not(location)/>
```

4.4 Other examples

To show the range of metaphorical gestures that our model can generate, let's study how the general representation and processes we just defined can be used to express an enumeration and a contrast.

Enumerating a set An enumeration is a rhetorical structure that consists in denoting each element of a set. The elements are grounded as objects inside a container and the enumeration operation consists in depicting the "contains" property of the container. This creates a unit with one discourse root whose goal is to enumerate each elements. Each element inside the container is then sequentially depicted at the container's location.

```
(operation (depict_property(c contains)))
--> <goals goal=enumerate type=container id=GenId() location=(location c)/>
    for each element (contains c e):
      <goal=depict location=(location c) type=object/>
    </goals>
```

Example At the beginning of the conversation, the clinician counts on her left-hand fingers the specific facts mentioned by the patient's husband. (see Figure 4).

Contrast A contrast shows the difference of value between a property of two elements. For example, we compare the size of two persons or the duration of two movies.

When applied to the concept, the property returns a scalar value that is grounded as a location on an axis (horizontal, vertical or frontal axis). Most of the properties returning a scalar value use the vertical axis (following the metaphor More is Up). Properties associated to time points use the horizontal axis when the speaker is not actively involved (Progression is a Writing Line), and the frontal line when she is (Progression is Moving Forward) [2].

Example When the clinician says: "You and him feel a little bit distant compared to how it used to be in the past.", she contrasts the situation now and how it was. Her mental space is described on the first column above.

To ground this mental space, the Grounded Conceptualizer first retrieves the axis associated to the *time* property (*horizontal axis*). Then, the abstract values (*now* and *past*) returned by the p*time* roperty are mapped to locations on the axis; the primary metaphors Past is Left and Present is Center are used. Finally, the communicative intention is mapped to depicting the two locations, leading to a sequence of gestures that locates each object (see Figures 5 and 6).

Mental Space	*Grounded Mental Space*
`(type s1 Situation)`	
`(actors s1 (patient husband))`	
`(time s1 now)`	`(operation depict_property(location s1 s2))`
`(type s2 Situation)`	`(type isa Object)`
`(actors s2 (patient husband))`	`(location s1 horizontal:center)`
`(time s2 past)`	`(type s2 Object)`
`(operation contrast(time s1 s2))`	`(location s2 horizontal:left)`

5 Discussion

In this paper, we showed that our model can transform various different mental spaces into gestures specifications by grounding their elements in a physical context. We particularly detailed the conceptual mapping from *abstract object* and *set* to *concrete object* and *container*. We showed how properties in abstract propositions can be conveyed by physical properties of gestures and presented how a relatively small set of operations can combine the physical components of the grounded mental spaces to convey a speaker's more elaborated communicative intention such as an enumeration or a contrast over time.

Even though the model presented in this paper focuses on the generation of metaphorical gestures, we believe that it is generic enough to take into account other kinds of gestures, namely *iconics* and *deictics*. Since *Iconics* consist in depicting one physical property of an object, the mental space is already grounded and the gesture can be directed specified. *Deictics* require information about the objects location in the physical space so pointing gestures can be appropriately generated. Therefore, integrating these other gestures in this framework seem, *a priori*, feasible.

Another question regards the generation of multimodal performances that would go beyond gestures. Two options can be considered. First, this model could be coupled to a behavior planner that generates nonverbal behaviors using either a natural language generator capable of also generating communicative intentions along with the utterance, or an inference process that would derive the underlying mental space from the utterance text and audio (in a process similar to inference-based behavior planners like Cerebella [17]). Generating nonverbal behaviors based on what is expressed in the speech confines the gesture performance as an illustration of the speech. The other -preferred- option consists in integrating this model in an architecture that generates speech and gesture. One lead is the model of [5] because it supports an arbitrary granularity of semantics that would allow us to align the natural language generation to the granularity of our mental space representation.

A key issue with the current implementation concerns selecting between alternative metaphors. For example, to convey that one idea is important, the grounding conceptualizer detects that “idea” and “importance” are abstract concepts and try to map them to concrete objects and properties. Therefore, it has to select which property is appropriate to represent the abstract notion of importance. By using primary metaphors, either the size or the weight could convey the desired intent[6]. Currently, we randomly pick one candidate mapping. Other options would be to internally evaluate the performances resulting from all candidate mappings. Because conflicts may arise downstream, such as incompatible grounded spaces or the absence of an appropriate gesture in the virtual human repertoire, the generated performance might not actually convey the speaker’s intentions. Preferences could be propagated backwards to decide which mapping to use in this specific context.

A preferable option would be to have a deeper model of what influences this choice. Several researchers have focused on this question. For example, the Structure Mapping Theory computes similarity by detecting a similar structure in source and target elements [7]. Since primary metaphors are not based on any objective similarity, they cannot be detected or generated by similarity-based model.

More fundamentally, this problem arises because we currently separate the reasoning and thought process from the conceptual grounding process. In a more radical view of embodied cognition, they could be treated as a combined process and the mental spaces would be inherently grounded. The grounding conceptualizer would not be a process in itself, but an underlying process on top of which our whole thinking and reasoning system is based. Creating such a model would require to understand -and computationally model- the whole range and dynamics of thought and reasoning, which still seems quite unrealistic.

Instead, we propose to keep on incrementally leveraging the mental space representation and its relation to gesture, speech and nonverbal behaviors. We will study what information is salient in speech, other kinds of gestures and nonverbal behaviors, and how these modalities relate to each other. Each one can embellish, substitute for and even contradict the information conveyed by the others [6]. For example, the same communicative intention can be expressed by co-occurrent yet different metaphors in speech and gesture [4]. This raises a number of issues concerning whether each modality can have its own grounded space, what determines what part is conveyed by each modality and how these modalities are synchronized.

[6] Much more information about which salient property to express is required to guide the grounding process and address the subtle distinction that a “heavy” decision has negative outcomes if one is wrong.

Beyond providing virtual humans with better communication skills, addressing such questions will inform both the representation of the mental space and the dynamic processes resulting from thought and reasoning, from a perspective that merges cognitive linguistics and gesture studies.

References

1. Bergmann, K., Kahl, S., Kopp, S.: Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In: Intelligent Virtual Agents. p. 203–216 (2013)
2. Calbris, G.: From left to right: Coverbal gestures and their symbolic use of space. Metaphor and gesture p. 27–53 (2008)
3. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: Proc.of the 28th conference on Computer graphics and interactive techniques. p. 477–486. SIGGRAPH '01, ACM, New York, USA (2001)
4. Cienki, A., Müller, C.: Metaphor and gesture. John Benjamins Pub. Co. (2008)
5. DeVault, D., Traum, D., Artstein, R.: Practical grammar-based NLG from examples. In: Proc. of the 5th International Natural Language Generation Conference. p. 77–85. Asso. for Computational Linguistics (2008)
6. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry: Journal for the Study of Interpersonal Processes 32(1), 88–106 (1969)
7. Gentner, D.: Structure-Mapping: a theoretical framework for analogy. Cognitive Science 7(2), 155–170 (Apr 1983)
8. Grady, J., Oakley, T., Coulson, S.: Blending and metaphor. Metaphor in Cognitive Linguistics p. 101–124 (1999)
9. Kendon, A.: Language and gesture. In: McNeill, D. (ed.) Language and gesture, p. 47–63. No. 2 in Language, culture & cognition, Cambridge Univ. P. (2000)
10. Kopp, S., Bergmann, K.: Individualized gesture production in embodied conversational agents. Human-Computer Interaction p. 287–301 (2012)
11. Kopp, S., Wachsmuth, I.: Model-based animation of co-verbal gesture. In: Proceedings of Computer Animation. p. 252–257 (2002)
12. Lakoff, G., Johnson, M.: Metaphors we live by. Univ of Chicago Press (1980)
13. Lascarides, A., Stone, M.: A formal semantic analysis of gesture. Journal of Semantics 26(4), 393–449 (Nov 2009)
14. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Conference on Autonomous Agents and Multiagent Systems. p. 289–296 (2009)
15. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Intelligent Virtual Agents (2006)
16. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. ACM Trans. Graph. 29(4), 124:1–124:11 (2010)
17. Lhommet, M., Marsella, S.: Gesture with meaning. In: Intelligent Virtual Agents, pp. 303–312 (2013)
18. McNeill, D.: Gesture and thought. Univ. of Chicago Press (2005)
19. Neff, M., Kipp, M., Albrecht, I., Seidel, H.: Gesture modeling and animation based on a probabilistic recreation of speaker style. ACM Transactions on Graphics 27(1), 5 (2008)