Gesture with meaning

Margaux Lhommet and Stacy Marsella

USC Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista, CA
{lhommet,marsella}@ict.usc.edu

Abstract. Embodied conversational agents (ECA) should exhibit nonverbal behaviors that are meaningfully related to their speech and mental state. This paper describes Cerebella, a system that automatically derives communicative functions from the text and audio of an utterance by combining lexical, acoustic, syntactic, semantic and rhetorical analyses. Communicative functions are then mapped to a multimodal behavior performance. Two studies demonstrate that the generated performances are meaningful and consistent with the speech.

Keywords: nonverbal behavior, embodied conversational agent

1 Introduction

Although it may seem that minds somehow directly interact, human face-to-face interaction is realized through the body. Beyond the words uttered, nonverbal behavior such as the flip of a hand, a gaze aversion, or a slumped posture, can powerfully influence interaction. These behaviors are so pervasive in every moment of the dialog that their absence also signals information - that something is wrong, for example, about the physical health or mental state of the person.

Our interest in such behaviors lies in a desire to model and automate the generation of nonverbal behavior for convincing, life-like virtual character performances.

A key challenge to the automation is understanding the nature of this nonverbal channel. Nonverbal behaviors establish a pervasive flow of information between participants in a conversation, because there is a rich interconnection between a person's mental processes and their body. Communicative intentions are conveyed, providing information that embellishes, substitutes for and even contradicts the information provided verbally (e.g., [1, 2]). Shifts in topic can be cued by shifts in posture or shifts in head pose. Comparison and contrasts between abstract ideas can be emphasized by abstract deictic (pointing) gestures that point at the opposing ideas as if they each had a distinct physical locus in space [3]. The form of these behaviors is often tied to physical metaphors thus underscoring the close connection between mental processes and the body. For example, the rejection of an idea can be illustrated by a sideways flip of the hand that suggests discarding an object as if an idea was a physical object [4]. Nonverbal behavior is also a reflection of the speaker's mental state. Gaze reveals thought processes, blushing suggests shyness and facial expressions, unintentionally or intentionally, convey emotions and attitudes.

The focus of our work is on automatic approaches to generate expressive, life-like nonverbal behavior. We have developed a flexible technique that employs information about the character's mental state and communicative intent to generate nonverbal behavior when that information is available. Otherwise, it uses acoustic, syntactic, semantic, pragmatic and rhetorical analyses of the utterance text and audio to infer the communicative functions (CFs). This includes deriving both the communicative intent of the utterance as well as the underlying emotional and mental state of the speaker. In either case, the CFs are then mapped to nonverbal behaviors, including head movements, facial expressions, gaze and gestures, that are composed and co-articulated into a final performance by a character animation system. In this paper, we give a broad overview of the approach and detail the rhetorical and semantic analyses that detect the CFs. We then report on two evaluation studies using human subjects that assess the consistency of generated performances with the speech.

2 Related Work

Researchers have explored techniques to generate nonverbal behavior, differing in how the models were developed, the degree of automation in the generation process itself and the particular classes of nonverbal behaviors that are handled.

Utterances can be manually annotated to specify what information has to be conveyed nonverbally. Annotations are then automatically mapped to appropriate nonverbal behaviors (e.g. [5, 6]). They are also used to communicate knowledge about the character's personality [7] or relationships [8].

Researchers have explored fully automatic generation of specific classes of nonverbal behaviors, using data-driven techniques. This includes models that generate gestures [9] or head movements [10] just by considering prosody, models that learn the mapping between speech text and head movements [11], and models of how speakers' gesture style differ [12, 13].

Also, there is work on nonverbal behavior generation using manually constructed models. BEAT automatically generates the speech and associated nonverbal behavior given the text of the utterance and infers rheme and theme to determine intonation and emphasis [14]. The NonVerbal Behavior Generator (NVBG) [15] extends this analysis by inferring the CFs embedded in the surface text (e.g. affirmation, intensification, negation, disfluencies) by using a keywords mapping. When integrated into a larger virtual human architecture, NVBG automatically associates CFs to provided information (emotional state, coping strategy and dialog acts).

BEAT and NVBG can be viewed as the intellectual ancestors to Cerebella. However, the limited analyses that drive those systems also limit both the nature of CFs detected as well as the frequency of detection. The central contribution of this work is the integration of a wider range of analyses, such as acoustic, rhetorical and semantic analyses of the text and audio of the utterance. This leads to a richer display of behaviors that are more meaningfully related to the utterance.

3 System Overview

Cerebella follows the SAIBA framework guidelines[1]. It takes as input communicative intents and generates a multimodal realization of this intent using the Behavior Markup Language (BML) [16], a high-level XML language that describes a behavior and an execution schedule and provides an abstraction to the animation system. Our system does not make strong assumptions about the provided inputs. If a complete Function Markup Language (FML) input containing detailed information about the mental state and communicative intents is provided, a direct mapping to nonverbal behaviors can be made. However, when only the utterance text and/or audio are given, the system tries to infer the communicative functions (CFs) through several analyses of the speech. We focus here on this last case.

Figure 1 presents an overview of our rule-based system. The central element is the Working Memory (WM) that stores the knowledge of the virtual human. The processing pipeline contains four sequential processes, detailed below.

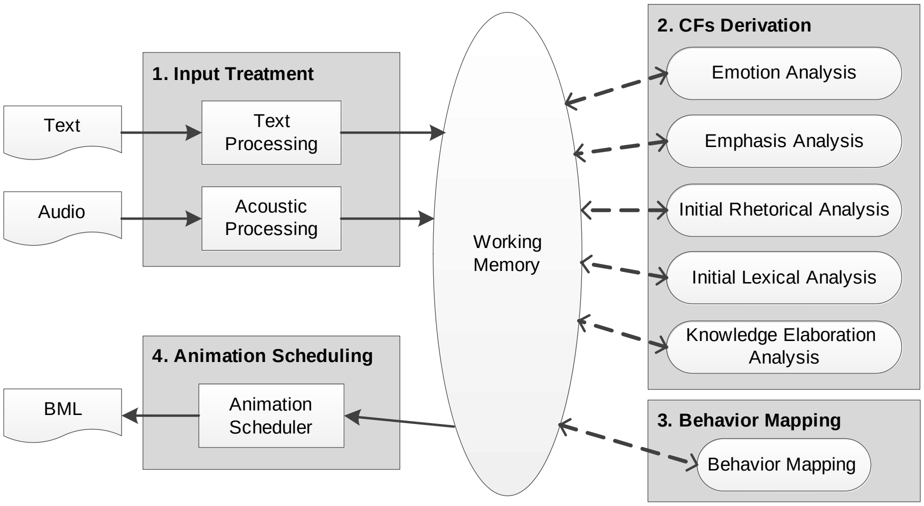

Fig. 1. Overview of Cerebella's processing pipeline when inferring communicative functions from text and audio.

3.1 Input Treatment

The input text is tokenized and each token is added to the WM. A natural language parser derives the syntactic structure [17], and each element of the

[1] http://www.mindmakers.org/projects/saiba/wiki

resulting parse tree is added to the WM. A limitation encountered is that most parsers are destined to text and not to partial utterances with their disfluencies and non-grammatical constructions.

An acoustic pipeline processes the audio of the spoken utterance. Two elements are currently derived by our system, overall agitation and word stress, relying on [18].

3.2 Communicative Functions Derivation

This phase detects the CFs present in the utterance. Each analysis consists of rules that match the content of the WM to infer new knowledge. The rules run in parallel so a rule can exploit knowledge inferred by the others. First, some rules take care of the low-level analyses (Emphasis, Emotion, Initial Lexical and Initial Rhetorical Analyses). Then the Knowledge Elaboration Analysis combines the inferred knowledge, leveraging the CFs detected. When no more rules match the content of the WM, the derivation phase ends. Table 1 shows the CFs that Cerebella currently derives, grouped into categories.

Table 1. Communicative Functions

Communicative Function Group	Communicative Function
-	Interrogative, negation, affirmation, emphasis
Rhetorical	Contrast, enumeration, alternative, accumulation, comparison
Intensifier	Strong positive, weak positive, strong negative, weak negative
Quantifier	Nothing, few, many, all, over, approximation
Comparative	Positive, negative
Time	Now, before, after, period
Location	Here, away
Deixis	You, me, we, abstract left, abstract right
Mental state	Cognitive load, emotional states

Emphasis Analysis: uses the word stress knowledge inferred by the Acoustic processing module to detect which words of the sentence are emphasized.

Emotion Analysis: uses the overall agitation level detected by the Acoustic processing module to determine the emotional arousal of the virtual human. We associate tense speech to high arousal, modal speech to mid-level arousal, and lax speech to low arousal.

Initial Lexical Analysis: the initial lexical analysis serves two purposes:

Lexical classification: maps a list of words and phrases to CFs, such as a *deixis_you* with words like “you” or “yourself”, or a *quantifier_nothing* with the words “nobody”, “none”, “nothing” or “never”. In addition, the WordNet database [19] is used to expand this knowledge base. In this case, a *quantifier_nothing* is detected whenever the surface text word is hierarchically linked to one of the associated WordNet synsets (“nothing.n.01”, “emptiness.n.01”).

Abstract/concrete classification: metaphors in language are often linked to metaphorical gestures, for instance when referring to an idea as if it were a concrete object, allowing to discard it with the flip of a hand [3]. To assist in this analysis, the system uses WordNet to annotate whether the nouns of the sentence refer to concrete or abstract concepts.

Initial Rhetorical Analysis: this analysis detects the CFs related to the rhetorical structure of the sentence. We rely on two assumptions to perform this analysis. First, rhetorical relations can be computed without a complete semantic analysis of the sentences (e.g. [20]). Second, such structural analyses work well enough at the utterance level to support nonverbal generation. The rules only use the knowledge asserted by the Text Processing module to detect the rhetorical constructions and their associated CFs. We detail here some of the rules that detect contrast, comparative and comparison. Similar rules allow detecting other rhetorical constructions such as enumeration, addition and alternative.

Comparative and comparison: a *comparative_positive* function is associated to the detection of "more" followed by a noun phrase, or to adjectives syntactically tagged as comparatives (e.g. "better", "smaller", "easier"). Moreover, whenever a *comparative* is followed by the word "than", a *rhetorical_comparison* is detected. For example, the detection of a *comparative_positive* on the beginning of the phrase "more interesting than before" will further lead to the association of a *rhetorical_comparison* to the whole phrase.

Contrast: a *rhetorical_contrast* function is detected, for example, by the following rule: word/expressions such as "but", "however", "unlike", ... surrounded by two part-of-speech that belong to the same syntactic category. When processing the sentence "This is really more interesting than before but I can only afford around 50 dollars", the previous rule will match and detect a *rhetorical_contrast* between the beginning of the sentence ("This is really more interesting than before") and the end ("I can only afford around 50 dollars").

Knowledge Elaboration Analysis: tests the knowledge previously asserted by the different analyzers and deepens, alters or removes the knowledge whenever required. We identify four purposes:

Combining CFs together requires verifying that the global meaning is coherent, since the CF of separate elements may be completely different when taken together. For example, in the expression "absolutely not", "absolutely" is associated to an *intensifier_positive* function (by the lexical classification process), and "not" to a *negative* function. The same issue can also be seen in more elaborated inferred knowledge. For example, in the expression "absolutely not more interesting", a *comparative_positive* function ("more interesting") and an *intensifier_negative* function are detected. That would generate separate and inappropriate gestures. The benefit of using a rule-base system is that it simplifies the combinatorics, e.g. positive or negative valence can be associated to different type of knowledge elements, and there are multiple ways of combining them. We do not want to explicitly enumerate all combinations but rather rely on rule-based forward-chaining.

Solving conflicts: it sometimes occurs that functions conflict with each other. For example, a *comparative* as well as an *emphasis* may be associated to the same word. So as part of the function derivation phase, each function inferred is assigned with a priority based on its CF and whether the words it spans are emphasized. Then these priorities are used to resolve conflicts between overlapping functions with lower priorities being dropped.

Semantic disambiguation: the initial lexical analysis is sometimes not sufficient to determine the actual meaning of a word or phrase. For example, the *quantifier_approximation* is associated to the word "about" each times it is encountered. However, this association can be wrong, such as in "I was thinking about you". Some rules, by combining semantic and syntactic information, help distinguish those cases, for example by testing that the element associated to the supposed *quantifier* is a number.

Expanding the CF: when detected, a CF is associated to the words matching the detection pattern of the rule. In some cases, the matched words may not span the full phrase that realizes that CF. For example, a *time_before* function is associated to the word "ago" in the expression "two years ago", but the function should cover the whole expression to generate synchronized gestures. Therefore, Cerebella contains a set of generic rules that span the CF over groups of words.

3.3 Behavior Mapping

The CFs derived during the previous phase are mapped to a set of alternative sequences of behavioral types to generate a schedule of multimodal behavior. For example, a *rhetorical_contrast* function might be realized by a synchronized tilting of the head and appropriate gesture. The alternatives allow variability in the character's behavior from one utterance to the next, as well as specialization by character. For example, the agitation state derived from the audio affects this mapping. Characters in the low agitation state (sad or lethargic) are biased to move heads from side to side instead of front to back. Highly agitated characters (angry or energetic) emphasize points using behaviors that include a beat rather than just subtler eyebrow raises.

3.4 Animation Scheduling

The multimodal nonverbal behaviors are mapped to the BML language. Behaviors that can be specified include head movements, gazing, blinking, saccadic eye movements, gesturing, facial expressions and speech. Behaviors are specified with start and end times such that they correspond to word starts or endings, or other behaviors when they are part of a sequence. Finally, the Smart Body animation system [21] interprets these high-level instructions to synthesize the final motion.

3.5 Knowledge

The knowledge used in Cerebella comes from diverse sources. The CFs derivation phase is currently driven by handcrafted rules and associated internal and external (specifically WordNet) databases, but more automatic approaches are being explored, such as using a learning-based rhetorical parser (SPADE [22]).

The knowledge used in the system represents a multi-year effort. Initially, an extensive literature review of the research on nonverbal behavior was undertaken. This initiated the design of rules encoding the function derivation and behavior mapping rules. Also, videos of real human face-to-face interactions have been annotated and analyzed to verify the rule knowledge. This annotation and analysis was critical because existing literature says little about dynamics of behaviors. We characterize this approach as a *expert knowledge plus semi-automated analysis* approach. More recently, pure data-driven machine learning techniques have been used as a way to validate the features used in the rules and to learn the mapping between features of an utterance and nonverbal behaviors [11].

4 Studies

The baseline hypothesis behind the inferencing that takes place during the CFs derivation process is that it will lead to gestures that will convey the meaning of the functions that are inferred. Here we conducted two studies to test this hypothesis.

4.1 First study

This study tests that the gestures generated by Cerebella convey the same meaning as the speech. We used 9 sentences containing CFs presented in this paper (time, comparative, location and quantity) to generate 9 virtual human video performances. The sound was removed from videos, leaving nonverbal behaviors as the sole indicator of meaning. 34 native English participants (14 female) completed the study via Amazon Mechanical Turk[2]. They could watch each video as many times as they wanted then had to select the sentence that matched the virtual human gestures in a set. Proposed choices included the original sentence as well as derivations created by reversing the original functional class(es). For example, the sentence “It is said that Spanish is much easier to learn than French” was derived into “It is said that French is less easy to learn than Spanish”. This sentence tests the association of the gesture (“a two-hand gap that increases”) to a *comparative_positive* function as opposed to a *comparative_negative* function. A choice of the original sentence implies a closer match between what the gesture and the original sentence convey, thereby helping to validate the CFs detection and behavior mapping used in our system.

Figure 2 shows the percentage of recognition of the functional classes. The red line marks the recognition rate that would be obtained by randomly selecting the answers. The participants were globally able to retrieve the original sentence by using the associate gesture (overall recognition percentage is above 50%). This is particularly true for most of the functional classes (with a recognition score between 67% and 85%), except when required to associate the classic “oscillating bowl” (described in [4]) to a *quantity_approximation* instead of a *quantity_few* (score=52.9%).

[2] http://aws.amazon.com/mturk/

4.2 Second study

This study evaluates the appropriateness of gestures regarding the content of the speech. We created 11 sets of 3 virtual human performances that were identical except for gestures. The first performance was accompanied by gestures generated by Cerebella (appropriate condition). The second one used gestures conveying the opposite meaning of the sentence and was generated by reversing the functional intents detected in the appropriate performance (opposite condition). The third one replaced the appropriate gestures by randomly selected gestures (random condition). 46 (26 female) native English speakers completed the study via Amazon Mechanical Turk. They had to watch the 3 videos and order them from 1 ("most consistent with the speech") to 3 ("less consistent with the speech").

Figure 3 shows the frequency with which each performance was ranked as the first choice. Performances generated by Cerebella are rated as the more consistent with the speech (f=0.57), followed by the random ones (f=0.24) and the opposite ones (f=0.19). A one-way ANOVA was conducted to compare the group effects between the different performances. Across the group a significant effect could be observed (F(2,135) =79.96, p<.0001). Post-hoc comparisons using the Tukey HSD (p<.01) indicate that Cerebella's performance frequency is significantly higher than the two other ones, but no significant difference can be observed between the random and opposite conditions. However, the opposite performance is significantly the most frequently rated as second choice (f=0.43, p<0.01).

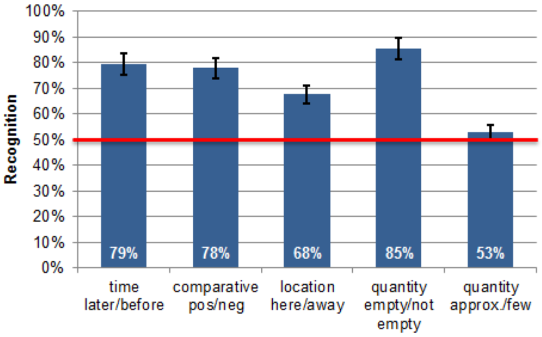

Fig. 2. First study results

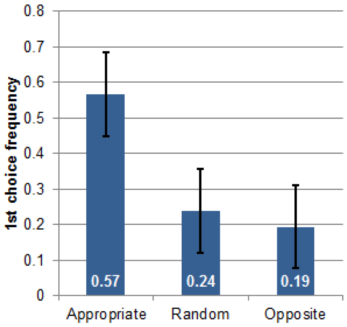

Fig. 3. Second study results

5 Discussion

Cerebella is an automatic approach to generate expressive, life-like nonverbal behavior. When available, it uses information about the character's CFs, including mental states and communicative intent, to generate behavior. Otherwise, it tries to infer CFs that underlie the input text and audio. Our system builds on previous works' approach [14, 15]. Acoustic, syntactic, semantic and rhetorical analyses of the utterance are designed to expand the CFs that can be detected as well as improve the accuracy of this detection.

As noted before, nonverbal behaviors express meaning through their form and dynamics. By inferring and exploiting this CF, generated nonverbal behavior ideally reflects that CF and even convey it by themselves. This paper presented two studies to corroborate this statement. Beyond this baseline hypothesis tested in this work, we plan to go on to assess whether the virtual human's gestures influence relational and cognitive factors including attitudes about the speaker, persuasiveness and recall.

While deeper and more elaborate analyses allow inferring and conveying CFs present in the sentence text and audio, this method encounters a particular limitation. As noted previously, nonverbal behavior can stand in a range of relations to the dialog. Here, the automatic generation of nonverbal behavior is limited in the range of CFs that can be inferred from the speech utterance only. This limitation is shared by all techniques that aim at automatically generating nonverbal behavior using speech. This can be overcome whether by taking as input those additional CFs or whether by integrating complex cognitive processes to generate them.

Moving forward with the system itself, one of the key issues will be to maintain a model of the relation between the CFs detected over the course of an interaction. For example, a speaker's gestures can physically locate the elements of the discourse and use deictics to refer back to them [4]. Additionally, nonverbal behaviors are determined by culture, gender, personality, attitudes as well as the context in which the communication takes place [23]. Fortunately, the use of a rule-based system combined to the staged approach we have taken will allow us to easily integrate new sources of input and broaden the range of CFs inferred and conveyed.

References

1. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica **1** (1969) 49–98
2. Kendon, A.: Language and gesture: Unity or duality. In McNeill, D., ed.: Language and gesture. Number 2 in Language, culture & cognition. Cambridge University Press (2000) 4763
3. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
4. Calbris, G.: Elements of Meaning in Gesture. John Benjamins Publishing (November 2011)
5. Kopp, S., Wachsmuth, I.: Model-based animation of co-verbal gesture. In: Computer Animation, 2002. Proceedings of. (2002) 252257
6. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with hands: creating animated conversational characters from recordings of human performance. In: ACM SIGGRAPH 2004 Papers. SIGGRAPH '04, New York, NY, USA, ACM (2004) 506513

7. Mancini, M., Pelachaud, C.: Generating distinctive behavior for embodied conversational agents. Journal on Multimodal User Interfaces **3**(4) (2009) 249261
8. Bickmore, T.: Relational Agents: Effecting Change through Human-Computer Relationships. PhD thesis, Massachusetts Institute of Technology (2003)
9. Levine, S., Krhenbhl, P., Thrun, S., Koltun, V.: Gesture controllers. In: ACM SIGGRAPH 2010 papers. SIGGRAPH '10, New York, NY, USA, ACM (2010) 124:1124:11
10. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. Audio, Speech, and Language Processing, IEEE Transactions on **15**(3) (2007) 10751086
11. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. (2009) 289296
12. Kopp, S., Bergmann, K.: Individualized gesture production in embodied conversational agents. Human-Computer Interaction: The Agency Perspective (2012) 287301
13. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) **27**(1) (2008) 5
14. Cassell, J., Vilhjlmsson, H.H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. (2001) 477486
15. Lee, J., Marsella, S.C.: Nonverbal behavior generator for embodied conversational agents. In: 6th International Conference on Intelligent Virtual Agents, Marina del Rey, CA (August 2006)
16. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thrisson, K., Vilhjlmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: Intelligent Virtual Agents. (2006) 205217
17. Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. (2000) 132139
18. Scherer, S., Kane, J., Gobl, C., Schwenker, F.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Computer Speech and Language **27**(1) (2013) 263–287
19. Miller, G.A.: WordNet: a lexical database for english. Communications of the ACM **38**(11) (1995) 3941
20. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press (2000)
21. Thiebaux, M., Marsella, S., Marshall, A.N., Kallmann, M.: SmartBody: behavior realization for embodied conversational agents. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1. AAMAS '08, Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems (2008) 151158
22. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. (2003) 149156