

Modeling and evaluating a Bayesian network of culture-dependent behaviors

Birgit Lugin, Julian Frommel, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Lugin, Birgit, Julian Frommel, and Elisabeth André. 2015. "Modeling and evaluating a Bayesian network of culture-dependent behaviors." In *2015 International Conference on Culture and Computing (Culture Computing), 17-19 October 2015, Kyoto, Japan*, edited by Toru Ishida, Naoko Tosa, Kozaburo Hachimura, Donghui Lin, Akira Maeda, and Matthias Rauterberg, 33–40. Piscataway, NJ: IEEE.
<https://doi.org/10.1109/culture.and.computing.2015.30>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



Modeling and Evaluating a Bayesian Network of Culture-dependent Behaviors

Birgit Lugin

Human-Computer Interaction
Würzburg University, Germany
birgit.lugin@uni-wuerzburg.de

Julian Frommel

Institute of Media Informatics
Ulm University, Germany
julian.frommel@uni-ulm.de

Elisabeth André

Human Centered Multimedia
Augsburg University, Germany
andre@hcm-lab.de

Abstract—Anthropomorphic user interfaces such as virtual agents or humanoid robots aim on simulating believable human behavior. As human behavior is influenced by diversifying factors such as cultural background, research in anthropomorphic user interfaces considers culture background for their behavioral models as well.

This paper presents a hybrid approach of creating a culture-specific model of non-verbal behaviors for simulated dialogs based on both: theoretical knowledge and empirical data. Therefore, the structure and variables of a Bayesian network are designed based on models and theories from the social sciences, while its parameters are learned from a video corpus of German and Japanese conversations in first time meeting scenarios. To validate the model a 10-fold-cross-validation has been conducted, suggesting that with the model culture-specific behavior can automatically be generated for some of the investigated behavioral aspects.

I. MOTIVATION

Non-verbal behavior takes a significant role during interpersonal interaction, e.g. [1], [2]. How these non-verbal behaviors are conducted and perceived is, amongst others, dependent on cultural background [3]. Unintended messages might be perceived due to different cultural backgrounds of the interlocutors. An example includes the expressiveness of gestures. *“What might seem like violent gesticulating to someone from Japan would seem quite normal and usual to someone from a Latin culture”* [4].

In a similar manner, a virtual character’s expressions can be perceived differently due to the cultural background of the observer. Studies by Koda and colleagues [5], for example, show that observers from different cultures judge the emotions of a virtual character differently based on its facial expressions. Therefore designers of virtual characters should take potential cultural differences into account when simulating natural non-verbal behaviors with virtual characters. This localization of the characters’ behaviors is likely to improve their acceptance by users of the targeted cultures. Vice versa, these characters can be used to show and explain cultural differences to users of a different cultural background.

Building models that determine culture-related differences in behavior is challenging as the causal relation of culture and corresponding behavior needs to be simulated in a convincing and consistent manner. In this paper, for the first time we provide such a model that is based on theoretical knowledge and augmented with empirical data using methods and algorithms from artificial intelligence. Therefore, we built

a Bayesian network, where cultural background and verbal behavior is used as causes and the resulting non-verbal behavior is calculated as effects. Applying an approach using Bayesian networks appears well suited, as they allow dealing with uncertain knowledge resulting from the fact that there is no deterministic mapping between human factors such as cultural background and non-verbal behavior. In addition, culture as a nondeterministic concept can be modeled without giving up a certain amount of variability that is necessary to ensure that an agent is perceived as an individual. Similar attempts have been made, for instance, for the generation of gestures supporting spatial information [6]. Learning the parameters of a Bayesian network from a multi-modal corpus has not been applied to culture-specific behaviors yet.

The present paper contributes to the research area of culture and computing by providing a statistical model that generates culture-dependent behavior for anthropomorphic user interfaces, as well as an evaluation of the model. Besides being able to use such a model to automatically generate culture-specific non-verbal behaviors for artificial dialogs, having such a model at hand allows further investigation of culture-related dependencies between the investigated behavioral cues that might stay unnoticed without a computational model.

II. RELATED WORK

The majority of approaches that investigate culture for anthropomorphic user interfaces is theory-driven. A common way of implementing the theory-driven approach is to start from existing multi-agent architectures and extend them to allow for culture-specific adaption of goals, beliefs and plans. One of the earliest and most well-known systems is the Tactical Language Training System (TLTS) [7] which is based on an architecture that implements a version of theory of mind, and has formed the basis of a variety of products for language and culture training. More recently systems have been developed that extend the agent mind architecture FAtiMA [8] to model rituals of the agents’ culture. Using this architecture, the ORIENT [9], the MIXER [10] and the Traveller [11] applications simulate abstract cultures that are based on theoretical knowledge with the overall aim to generate more cultural awareness on the user’s side.

Data-driven approaches, on the other hand, use data such as annotated multimodal recordings of existing cultures as a basis for computational models of culture. Such a cross-cultural corpus has, for example, been recorded for multi-party multi-modal dialogues in the Arab, American English and Mexican

Spanish cultures [12]. The corpus has been coded regarding proxemics, gaze and turn taking behaviors to allow extraction of culture-related differences in multi-party conversations.

While the theory-driven approach ensures a higher level of consistency than the data-driven approach, it is not grounded in empirical data and thus may not faithfully reflect the non-verbal behavior of existing cultures. Another limitation is that it is difficult to decide which non-verbal behaviors to choose for externalizing the goals and needs generated in the agent minds. The advantage of data-driven computational models of culture lies in their empirical foundation. However, they are hard to adapt to settings different from the ones recorded, as the data cannot be generalized for a lack of a causal model.

The hybrid approach, presented in this contribution, combines advantages of the theory-driven and data-driven approaches, as it explains the causal relations of cultural background and resulting behavior, and augments them by findings from empirical data.

III. BACKGROUND

In our own former work, we have been applying both approaches described above separately. In scope of the Cube-G project a multi-cultural corpus was recorded in the German and Japanese cultures [13]. More than 20 participants were recorded in each culture, each running through three scenarios. For this paper, the first scenario, a first time meeting, was considered. For this scenario a student and a professional actor (acting as another student) were told to get acquainted with one another to be able to solve a task together later. Recordings started during this conversation already. Recording one participant and an actor at a time ensured a higher control over the recordings. This way, participants did not know each other in advance and we the actor was able to control that the conversation lasted for five minutes. Actors were told to be as passive as possible to allow the participant to lead the conversation and be active in cases where the conversation was going to stagnate.

Statistical analyses of the video corpus were performed highlighting differences between the recorded cultures in both verbal [14] and non-verbal behavior [15]. At a later stage, we conducted perception studies with virtual characters that simulated the findings of the corpus analysis. Results suggest that users prefer virtual character behavior that was designed to resemble their own cultural background [16]. Please note, that in these studies, the characters' behavior was completely scripted to follow the statistical distribution of the corpus findings, and no computational model was built yet, while each of the studies looked at one behavioral aspect in isolation.

In parallel, to build a first computational model, we built a Bayesian network focusing on non-verbal behavior based on theoretical knowledge [15] without enhancing it with the corpus data yet.

Thus, the present contribution applies a hybrid approach using machine learning for the first time to augment a theoretical model with empirical data, including a validation of the model. With the model, we are now able to generate culture-specific dialog behavior automatically following the statistical distribution of the recorded data. In addition this contribution

combines for the first time all previously considered behavioral aspects (verbal and non-verbal) in a complete model, allowing further investigations of dependencies between culture-specific behavioral cues.

IV. NETWORK MODEL

The structure of the network with its variables was modeled based on cultural theories and categorizations of behavioral aspects and implemented using the GeNIe modeling environment [17]. Our aim is to generate non-verbal behavior for simulated dialogs in the domain of first time meetings. Therefore, the network is divided in two parts: influencing factors and (resulting) non-verbal behavior (see Figure 1).

A. Influencing Factors

In line with the objective of our work, influencing factors in our model are cultural background as well as verbal behavior.

To model **culture** in our network, we use Hofstede's dimensional model [18], which is very well suited for the implementation of computational models and has widely been used to simulate culture for anthropomorphic interfaces, e.g. [9], [10], [11]. For the model more than 70 cultures were categorized along 5 dimensions in an empirical survey. The *Power Distance* dimension (PDI) describes the extent to which a different distribution of power is accepted by the less powerful members of a culture. The *Individualism* dimension (IDV) describes the degree to which individuals are integrated into a group. On the individualist side ties between individuals are loose, while on the collectivist side, people are integrated into strong, cohesive in-groups. The *Masculinity* dimension (MAS) describes the distribution of roles between the genders. In feminine cultures, roles differ less than in masculine cultures, while competition is rather accepted in masculine cultures where status symbols are of importance. The *Uncertainty Avoidance* dimension (UAI) defines the tolerance for uncertainty and ambiguity. It indicates to what extent the members of a culture feel comfortable or uncomfortable in unstructured or unknown situations. The *Long-Term Orientation* dimension (LTO) explains differences in the perception of virtue. For each dimension, clear mappings are available from existing national cultures to the cultural dimensions on normalized scales between 0 and 100 [18]. In our network, we categorized the scores on the cultural dimensions into three discrete values (low, medium, high).

Dialog behavior was broken down to **speech acts** and **conversational topics** in our network. Speech acts can be categorized along the DAMSL (Dialog Act Markup in Several Layers) coding scheme that was introduced by Core and Allen [19]. One layer of the schema, labels the communicative meaning of a speech-act as needed to categorize our dialogs. According to our underlying scenario, a first time meeting, we use the following subset of communicative functions: *statement*, *answer*, *info request*, *agreement / disagreement* (indicating the speaker's point of view), *understanding / misunderstanding* (without stating a point of view), *hold*, *laugh* and *other*.

According to Schneider [20], topics that prototypically occur in first-time meetings can be classified as follows: The *immediate situation* holds topics that are elements of the so-called frame of the situation, such as the surrounding or

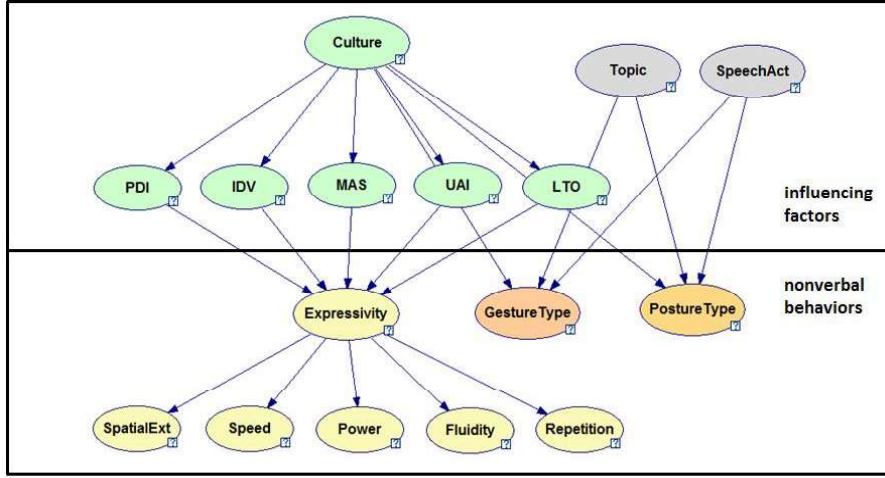


Fig. 1. Network model for culture-related non-verbal behavior generation.

the atmosphere of the conversation. The *external situation* describes all topics that hold the larger context of the immediate situation, such as the news, politics, sports or movies. For the *communication situation* interlocutors are seen as a subset of the immediate situation, with topics focusing on the conversation partners, e.g., their hobbies, family or career. Topics in our network are categorized accordingly.

B. Non-verbal Behavior

Regarding the resulting non-verbal behavior, we focus on gestures, gestural expressivity and postures. The generation of adequate body postures, gestures, and their expressivity have been widely studied in the field of virtual agents (e.g. [21], [22], [23]) and seem to be good aspects to improve the characters' believability. We aim on further enhancing them by adding a cultural perspective on these behavioral aspects. For the aspect of body posture, for example, it has been shown that different cultures perceive different emotions from body postures [24]. Other aspects such as eye gaze or head movements have not been taken into account yet in this paper.

To distinguish **posture types**, we employ Bull's posture categorization for arm postures [25]. In total, 32 different arm positions are presented in his categorization and included to the network, such as *PHEw* (put hands on elbow), *PHWr* (put hands on wrist) or *FAs* (fold arms). Please see [25] for a full list of arm postures.

We classify **gesture types** according to McNeill's categorization [26]: deictic, beat, emblem, iconic, metaphoric, and adaptors. *Deictic* gestures are pointing gestures. *Beat* gestures are rhythmic gestures that follow the prosody of speech. *Emblems* have a conventionalized meaning and do not need to be accompanied by speech. *Iconic* gestures explain the semantic content of speech, while *metaphoric* gestures accompany the semantic content of speech in an abstract manner by the use of metaphors. *Adaptors* are hand movements towards other parts of the body to satisfy bodily needs, such as scratching one's nose. As we are focusing on gestures that accompany speech, adaptors were excluded from the network.

Emblems were excluded as well, as they might convey different meanings in different locations. Thus, even if we could predict that a emblematic gesture type would be appropriate, different concrete gestures needed to be selected based on cultural background.

To describe the dynamics of a gesture, we added a node containing its **expressivity**, which can further be broken down into dimensions [27]. Following [28], who investigated gestural expressivity for virtual characters, we employ the parameters spatial extent, power, speed, fluidity and repetition: The **spatial extent** describes the arm's extent relative to the torso. The **speed** and **power** (acceleration) with which a gesture is performed can vary as well. The **fluidity** describes the flow of movements, as gestures can be conducted in a jerky or fluid manner. The **repetition** holds information about the repetition of the stroke of a gesture. In our network expressivity is categorized by the three values low, medium and high. Initial values were set in a manner that a high expressivity is more likely to result in a higher value for each of the parameters.

C. Dependencies

There is evidence from the literature that the choice of gesture types and posture types are dependent on cultural background, e.g. [3]. However there are no clear statements in the literature of how McNeill's gesture types or Bull's posture types would correlate with Hofstede's dimensions of culture. We thus connected the nodes holding gesture types and posture types directly to the culture node instead of linking them via Hofstede's dimensions.

Regarding non-verbal expressivity, prototypical behavioral traits depending on cultural dimensions are described in [29] where so-called synthetic cultures are introduced. Synthetic cultures describe fantasy cultures that find themselves on one extreme end of one of Hofstede's cultural dimensions. For example, the *extreme masculine culture* is described as being loud and verbal, liking physical contact, direct eye contact, and animated gestures. *Extreme feminine cultures*, on the other hand, do not raise their voices, like agreement, do not take

much room and are warm and friendly in conversations. The descriptions of prototypical behavior for synthetic cultures clearly indicate that the positioning on Hofstede’s dimensions has an impact on the level of expressivity. Thus, the dimensions are connected to the expressivity node.

It is known from the literature that verbal behavior, e.g. the choice of conversational topic, is amongst others dependent on cultural background. According to Isbister and colleagues [30], for example, the categorization into safe and unsafe topics varies with cultural background. However, the aim of our network model is to generate culture-dependent non-verbal behavior for a given dialog. We therefore do not yet investigate dependencies between nodes of our influencing factors (culture, speech act and conversational topic). However, the dialog behavior should influence the accompanying non-verbal behavior. Therefore, nodes holding information about the selected verbal behavior (speech act and conversational topic) are also directly linked to the nodes holding non-verbal behavior types.

V. PARAMETERS OF THE MODEL

To augment our network model with empirical data, the corpus findings had to be prepared for further processing and included to by an automated learning process.

The videos of the corpus were annotated for statistical analysis using the Anvil tool [31] that allows to specify attributes with their parameters and align them in a timely manner. Verbal behavior was annotated for conversational topics [20] and speech acts using the subset of the DAMSL coding scheme [19] mentioned above. For the annotation of postures, Bull’s posture coding scheme [25] was employed. Gestures were annotated according to McNeill’s gesture types [26] and the expressivity parameters [28]. For each participant the cultural background was added to the meta data of the annotations.

To use those annotations for a machine learning approach, in a first step, the different modalities needed to be aligned. Therefore, we divided each annotated conversation into conversational blocks that we further refer to as datasets. Depending on the annotations and with it the semantics of the speech, each dataset thus initially refers to a clause or a sub-clause. Each dataset is determined by this speech utterance, specified by its speech act and conversational topic.

Behavioral aspects occur at a certain time interval during the conversation. Based on the speech act, we added non-verbal behavior to the dataset, in case there is a overlap of their intervals. Thus, for each dataset, there may or may not be an accompanying posture and / or gesture available. In case a gesture or posture did not occur, an empty token is added to the dataset.

Gestures and postures are added to all speech acts that they overlapped with to reflect that a gesture or posture can be maintained for a longer time period. In case several gestures (or postures) overlapped with the same speech act, the gesture (or posture) is added to the dataset where the overlap with the speech act lasted for longer. Due to data loss, the timely information on the annotations of expressivity could not be aligned with the corresponding speech act, but only be used

quantitatively. To nevertheless be able to integrate the data to the network, two different datasets were used for the learning process. Firstly, we used the aligned dataset to learn the joint probability distributions of arm postures and gesture types (dependent on culture and verbal behavior). Secondly, we used the non-aligned dataset to learn the gestural expressivity (dependent on culture only). Therefore, the dependency of expressivity and verbal behavior was removed from our network model.

After the extraction, the aligned dataset contained a list of 2155 dataset values, and the non-aligned dataset contained 457 values. The SMILE-Framework underlying the GeNIe modeling environment [17] that was used to model the structure of our network, provides amongst others, an implementation of the EM-algorithm [32]. As not all aspects were annotated for each person recorded in the corpus, we have to deal with incomplete data. For example, there are some verbal annotations (speech act and topic) missing in the Japanese part of the corpus, as translation was not available. The EM-algorithm is thus well suited for our purpose, because it is capable of dealing with those incomplete datasets. In our two-folded learning approach, firstly, the aligned dataset containing information about the speech act was used to learn the probabilities of the parameters for posture and gesture types. Secondly, the non-aligned dataset was applied to determine the parameters for the gestural expressivity.

VI. RESULTING NETWORK

Figure 2 exemplifies calculations of the network with the evidence of cultural background being set to Japanese. As mentioned earlier, in our former statistical analysis, we have been looking at behavioral aspects depending on cultural background in isolation (e.g. [15]). In case only the evidence for cultural background is set, distributions reflect those findings. With this setting, cultural variation in non-verbal behavior can be reflected in a general manner based on culture only. For example, more expressive nonverbal behavior was observed in the prototypical German dialogs compared to the Japanese ones.

As the model contains several behavioral aspects, the learned Bayesian network additionally allows us to explore correlations in the data in an intuitive manner by setting additional evidences, e.g. for verbal behavior. For example, a correlation of chosen topic and non-verbal behavior frequency stayed unnoticed in our earlier work. From previous analysis of verbal behavior [14], we know that the topic distribution is different for the two cultures in the data. While in Japan significantly more topics covering the immediate situation occurred compared to Germany, in Germany significantly more topics covering the communication situation occurred compared to Japan. Setting evidences to the topic nodes, the network reveals that people in both cultures are more likely to perform gestures when talking about less common topics. In particular, the communication situation in the Japanese culture and the immediate situation in the German culture. This effect could be explained by the tendency that talking about a more uncommon topic might lead to a feeling of insecurity that results in an increased usage of gestures. Thus, the network also reveals how culture-related non-verbal is mediated by culture-specific variations in verbal behavior.

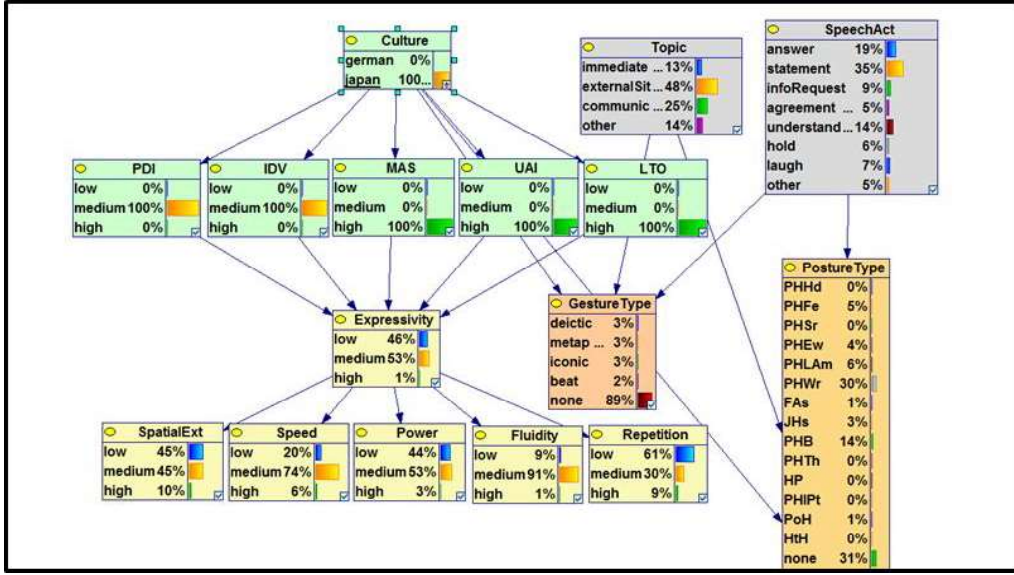


Fig. 2. Resulting Bayesian network including nodes and categories as described in section IV, as well as the parameters learned from the empirical data, with cultural background set to Japanese.

A. Demonstrator

As stated in section III, in our former work we performed perception studies with scripted behavior for virtual characters that follow the statistical analysis of our video corpus, using a virtual character system [33] containing culture-specific characters that match a prototypical Asian or Western ethnic background.

In comparison to the scripted perception studies, the Bayesian network is able to generate nonverbal behaviors for a given cultural background and a given agent dialog. For demonstration, a first time meeting dialog, similar to the ones recorded in the corpus, was tagged with categorizations of speech acts and topics. Probabilities for non-verbal behaviors are generated by the network depending on the current speech act, topic, and cultural background of the agents. In the demonstrator, postures and gestures are selected following the probability distribution of the network and is then simulated by the characters. Figure 3 shows a screenshot of a male German character in a prototypical German posture (FA - fold arms) in conversation with a female Japanese character performing an iconic gesture with a small spatial extent. With it, a certain variety in the characters' behaviors is preserved to present culture as a non-deterministic concept.

In our former studies scripted behavioral aspects were tested in isolation. Results suggested that observers tended to prefer virtual agent behavior that is in line with their own cultural background for some of the behavioral aspects [16]. Although an user evaluation with the demonstrator including the presented network was not performed yet, we hope to achieve similar results in case only the evidence for cultural background is set, as the virtual characters' non-verbal behavior generated by the network follows the same statistical distribution that was reflected in our former studies by the scripted behavior.



Fig. 3. Virtual characters showing prototypical culture-dependent non-verbal behaviors.

VII. EVALUATION

In order to validate the model a 10-fold-cross-validation was performed. The aligned dataset included 2155 dataset values where non-verbal behavior was aligned with the speech of the participants. Leaving every tenth dataset out, training the model with the remaining 90% and performing this step ten times, provides us with a validation set containing 2155 entries again. For each of these datasets, the cultural background (German or Japanese) is given, as well as the performed verbal behavior (speech act type and topic category). The accompanying non-verbal behavior (posture type and gesture type) is predicted by the network. Please note that we cannot validate the non-verbal expressivity using this approach due to the missing alignment. For all datasets the predictions of the network were tested against the behavior that was actually observed in the corpus data. In human behavior it appears quite unlikely for a person to perform the exact same way several times in a given situation. For a virtual character's behavior a

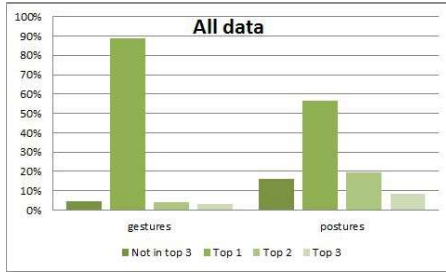


Fig. 4. Prediction rates of observed gesture types and posture types being in the first, second or third most likely gesture and posture type.

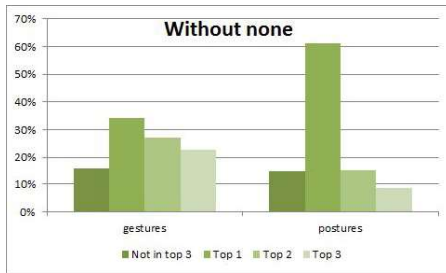


Fig. 5. Prediction rates of observed gesture type and posture type being in the first, second or third most likely gesture and posture type, excluding none-elements.

similar variety of behavior is desirable. With the overall aim of our project, to predict believable behavior for a given cultural background, we therefore additionally present in the following figures whether the performed gesture type or posture type is finding itself in the best three guesses of the network, or not.

Figure 4 shows the prediction rates for gesture types and posture types respectively. Results look quite promising, with an overall accuracy of 88% for gesture types and 56% for posture types. However, these results should not be overrated as for many of the observed speech acts no non-verbal behavior was conducted (resulting in the gesture and posture type "none"). In particular, only in 11% of our dataset a gesture was performed during a speech act, while in 71% of the speech acts a posture was performed.

We therefore excluded datasets where no gesture or posture was observed and performed another 10-fold-cross-validation leaving the option "none" out. With it, we evaluate the prediction rates of our network assuming to know that a gesture or posture should be performed by an agent. In total, 233 speech acts were accompanied by a gesture, 1551 by a posture. Figure 5 summarizes the results. Regarding gesture types, although a weak trend can be observed into the right direction, only 34% of the performed gestures were correctly predicted by the network which appears not better than random. The overall accuracy for posture types is 61% which looks much more promising.

In order to find out whether the predicted gesture and posture types reflect a prototypical cultural background, in a further evaluation step, we reversed the cultural background. Therefore, we performed a 10-fold-cross validation with the network being set to a Japanese cultural background for the

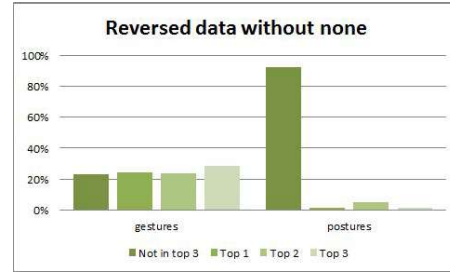


Fig. 6. Prediction rates of observed gesture types and posture types being in the first, second or third most likely gesture or posture type with the cultural background set to the reversed culture, excluding none-elements.

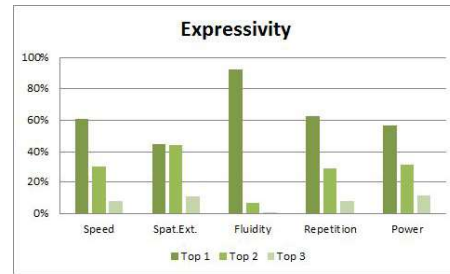


Fig. 7. Prediction rates of observed gestural expressivity being in the first, second or third most likely category.

German part of the validation set, and vice versa. Including none-elements, accuracy rate of gesture types is still 80%, as performing no gesture is the most likely prediction for both cultures. Leaving none-elements out, accuracy rate for gesture types drops to 24% (see Figure 6), resulting in worse predictions of the network compared to the original data set (cf. 5).

Regarding posture types, accuracy rate drops to 5% with the reversed cultural background including none-elements, with 75% of the observed postures falling not even in the top 3 categories. Excluding none-elements, the accuracy rate for postures with reversed cultural background is less than 2%, with 92% of the observed posture types not being in the top three guesses (see Figure 6). Thus, for posture types, changing the cultural background leads to a very low predictive power of the network, suggesting that the network is able to predict posture types dependent on the cultural background of the speaker.

Regarding expressivity we performed a 10-fold-cross validation for the parameters of expressivity based on cultural background only. We thus predict the probabilities for the level of expressivity given that a gesture is being performed. A more complete data set could be used in this case, as missing translations of verbal behavior could be ignored. In total 457 gestures were added to the validation set. Figure 7 shows the calculated levels of expressivity. Please note that in this case only three categories were available. Results look quite promising in this regards, suggesting that culture-dependent levels of expressive behavior can be predicted by trend, leaving very low prediction rated for the least likely category.

VIII. DISCUSSION

With the network, the most likely culture-specific non-verbal behavioral type (gesture and posture) is determined based on verbal behavior (speech act and topic) and cultural background. The most probable culture-related level of expressiveness is added based on cultural background. Regarding posture types and gestural expressivity, the presented network performed well. The strong correlation of cultural background and body posture in our data was also reflected by our previous statistical analysis, showing significant differences in the occurrence of posture types between the cultures. Regarding gestural expressivity, the data also revealed strong differences between the cultures. We thus believe that the network can help enculturating non-verbal behaviors in simulated dialogs for these aspects.

Regarding gesture-types, no reliable predictions could be made by our network based on culture. Thus, at the current stage, the network cannot add to believably simulating culture-specific behaviors focusing on gesture-types. This result is not surprising, considering the fact that our former statistical analysis showed that the overall number of gestures is similar in both cultures and no significant differences were found in the data regarding the frequencies of McNeill's gesture types. This might be caused by the abstraction of gestures to categories. Even if, for example, a deictic gesture is performed, the concrete execution should be different across cultures. While a deictic gesture is typically performed using the index finger in Western cultures, this is considered rude in some Asian cultures, where deictic gestures are usually performed using the whole hand.

IX. CONCLUSIONS AND FUTURE WORK

This paper presents a hybrid approach to model a Bayesian network that determines culture-dependent non-verbal dialog behavior for anthropomorphic user interfaces. In the hybrid approach, the structure of the network along with categorizations of behavioral aspects were constructed based on existing theories and models. The parameters of the network were learned from an annotated video corpus that recorded first-time meetings of German and Japanese participants respectively in scope of the Cube-G project [13]. The present contribution extends previous work by integrating aspects of verbal and nonverbal behavior into a complete model that can be used to automatically generate behavior that follows the statistical distributions of the underlying video corpus, while keeping a certain variability of behaviors. With the network, we are able to reflect cultural variation in non-verbal behavior based on culture, as well as simulate how culture-specific variations in verbal behavior mediate non-verbal behaviors.

The evaluation of the presented network shows promising results for some of the investigated behavioral aspects (postures and gestural expressivity), while it fails in predicting culture-related choices of gestures-types. We thus think that the network can be used to add posture-types and levels of expressive behavior to simulated dialogs in order to increase the culture-relatedness of the simulated non-verbal behaviors. For gesture-types further research is needed such as going into more depth regarding the performance of gestures or their correlation to the semantics of speech rather than speech acts.

In conclusion, the resulting network model is an approximation to reality in two ways: (1) the underlying theories and categorizations are in some cases too broad to serve as a basis for culturally dependent behavior generation, and (2) the resulting model can only be as meaningful as the data being used to specify the probabilities.

The model allows to be expanded by further aspects of culture-specific behaviors. In our future work, we aim on adding additional non-verbal behavioral traits that are known to be dependent on cultural background, such as head nods, to obtain a more complete model. In a similar way, a temporal component could improve the predictions of the network. For example, if a person performed a certain body posture during a speech act, it should be more likely that the same posture is performed during the subsequent speech act as well. We therefore aim on building a dynamic Bayesian network, that takes previously performed behaviors into account.

REFERENCES

- [1] K. Hogan, *Can't Get Through: Eight Barriers to Communication*. Pelican Publishing, 2003.
- [2] A. Mehrabian, *Silent messages: Implicit communication of emotions and attitudes*. Wadsworth, 1980.
- [3] S. Ting-Toomey, *Communicating across cultures*. New York: The Guilford Press, 1999.
- [4] K. Isbister, *Agent Culture: Human-Agent Interaction in a Multikultural World*. Lawrence Erlbaum Associates, 2004, ch. Building Bridges Through the Unspoken: Embodied Agents to Facilitate Intercultural Communication, pp. 233–244.
- [5] T. Koda, Z. Ruttkay, Y. Nakagawa, and K. Tabuchi, "Cross-Cultural Study on Facial Regions as Cues to Recognize Emotions of Virtual Agents," in *Culture and Computing*, T. Ishida, Ed. Springer, 2010, pp. 16–27.
- [6] K. Bergmann and S. Kopp, "Bayesian Decision Networks for Iconic Gesture Generation," in *Proc. of 9th Int. Conf. on Intelligent Virtual Agents (IVA 2009)*, Z. Ruttkay, M. Kipp, A. Nijholt, and H.-H. Vilhjálmsson, Eds. Springer, 2009, pp. 76–89.
- [7] W.-J. Johnson, S. Marsella, and H. Vilhjálmsson, "The DARWARS Tactical Language Training System," in *Interservice / Industry Training, Simulation, and Education Conference*, 2004.
- [8] J. Dias and A. Paiva, "Feeling and reasoning: a computational model," in *12th Portuguese Conference on Artificial Intelligence, EPIA*. Springer, 2005, pp. 127–140.
- [9] R. Aylett, A. Paiva, N. Vannini, S. Enz, E. André, and L. Hall, "But that was in another country: agents and intercultural empathy," in *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, 2009.
- [10] R. Aylett, L. Hall, S. Tazzymann, B. Endrass, E. André, C. Ritter, A. Nazir, A. Paiva, G. J. Hofstede, and A. Kappas, "Werewolves, Cheats, and Cultural Sensitivity," in *Proc. of 13th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, 2014.
- [11] S. Mascarenhas, A. Silva, A. Paiva, R. Aylett, F. Kistler, E. André, N. Deggens, G. J. Hofstede, and A. Kappas, "Traveller: an intercultural training system with intelligent agents," in *Proc. of 12th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, 2013.
- [12] D. Herrera, D. Novick, D. Jan, and D. R. Traum, "The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus," in *Multimodal Corpora Workshop: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*, 2010.
- [13] M. Rehm, E. André, Y. Nakano, T. Nishida, N. Bee, B. Endrass, H.-H. Huan, and M. Wissner, "The CUBE-G approach - Coaching culture-specific nonverbal behavior by virtual agents," in *ISAGA 2007: Organizing and Learning through Gaming and Simulation*, I. Mayer and H. Mastik, Eds., 2007.

- [14] B. Endrass, Y. Nakano, A. Lipi, M. Rehm, and E. André, "Culture-related topic selection in SmallTalk conversations across Germany and Japan," in *Proc. of 11th Int. Conf. on Intelligent Virtual Agents (IVA 2011)*, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, Eds. Springer, 2011, pp. 1–13.
- [15] M. Rehm, Y. Nakano, E. André, T. Nishida, N. Bee, B. Endrass, M. Wissner, A.-A. Lipi, and H.-H. Huang, "From observation to simulation: generating culture-specific behavior for interactive systems," *AI & Society*, vol. 24, no. 3, pp. 267–280, 2009.
- [16] B. Endrass, E. André, M. Rehm, and Y. Nakano, "Investigating culture-related aspects of behavior for virtual characters," *Autonomous Agents and Multi-Agent Systems*, 2013.
- [17] M. J. Druzdzel, "SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: A development environment for graphical decision-theoretic models (Intelligent Systems Demonstration)," in *Proc. of the 16th National Conf. on Artificial Intelligence (AAAI-99)*. AAAI Press, 1999, pp. 902–903.
- [18] G. Hofstede, G.-J. Hofstede, and M. Minkov, *Cultures and Organisations. SOFTWARE OF THE MIND. Intercultural Cooperation and its Importance for Survival*. McGraw Hill, 2010.
- [19] M. Core and J. Allen, "Coding Dialogs with the DAMSL Annotation Scheme," in *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, 1997, pp. 28–35.
- [20] K. P. Schneider, *Small Talk: Analysing Phatic Discourse*. Marburg: Hitzeroth, 1988.
- [21] J. Cassell, H. Vilhjálmsson, and T. Bickmore, "BEAT: The Behaviour Expression Animation Toolkit," in *Proc. of 28th Annual Conf. on Computer Graphics (SIGGRAPH 2001)*. ACM, 2001, pp. 477–486.
- [22] C. Pelachaud, "Multimodal expressive embodied conversational agents," in *Proc. of 13th annual ACM Int. Conf. on Multimedia*, 2005, pp. 683–689.
- [23] S. Buisine, M. Courgeon, A. Charles, C. Clavel, J.-C. Martin, N. Tan, and O. Grynszpan, "The role of body postures in the recognition of emotions in contextually rich scenarios," *International Journal of Human-Computer Interaction*, vol. 30, no. 1, 2014.
- [24] A. Kleinsmith, P. D. Silva, and N. Bianchi-Berthouze, "Recognizing emotion from postures: Cross-cultural differences in user modeling," in *User Modeling 2005*, ser. LNCS, no. 3538, 2005, pp. 50–59.
- [25] P. Bull, *Posture and Gesture*. Oxford: Pergamon Press, 1987.
- [26] D. McNeill, *Hand and Mind - What Gestures Reveal about Thought*. Chicago, London: University of Chicago Press, 1992.
- [27] P. E. Gallaher, "Individual Differences in Nonverbal Behavior: Dimensions of Style," *Journal of Personality and Social Psychology*, vol. 63, no. 1, pp. 133–145, 1992.
- [28] J.-C. Martin, S. Abrilian, L. Devillers, M. Lamolle, M. Mancini, and C. Pelachaud, "Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs," in *Proc. of 5th Int. Conf. on Intelligent Virtual Agents (IVA 2005)*. Springer, 2005, pp. 405–417.
- [29] G. J. Hofstede, P. B. Pedersen, and G. Hofstede, *Exploring Culture - Exercises, Stories and Synthetic Cultures*. Yarmouth, United States: Intercultural Press, 2002.
- [30] K. Isbister, H. Nakanishi, T. Ishida, and C. Nass, "Helper agent: Designing an assistant for human-human interaction in a virtual meeting space," in *Proc. of Int. Conf. on Human Factors in Computing Systems (CHI 2000)*, T. Turner and G. Szwillus, Eds. ACM, 2000, pp. 57–64.
- [31] M. Kipp, "Anvil - A Generic Annotation Tool for Multimodal Dialogue," in *Eurospeech 2001*, 2001, pp. 1367–1370.
- [32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
- [33] I. Damian, B. Endrass, P. Huber, N. Bee, and E. André, "Individualized Agent Interactions," in *Proc. of 4th Int. Conf. on Motion in Games (MIG 2011)*, 2011.