# Towards Natural Gesture Synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis

Michael Kipp[1], Michael Neff[2], Kerstin H. Kipp[3], and Irene Albrecht[4]

[1] DFKI, Germany, michael.kipp@dfki.de
[2] UC Davis, USA, neff@cs.ucdavis.edu
[3] Saarland University, Experimental Neuropsychology Unit, Germany, k.kipp@mx.uni-saarland.de
[4] TomTec Imaging Systems GmbH, Germany, ialbrecht@tomtec.de

**Abstract.** Virtual humans still lack naturalness in their nonverbal behaviour. We present a data-driven solution that moves towards a more natural synthesis of hand and arm gestures by recreating gestural behaviour in the style of a human performer. Our algorithm exploits the concept of gesture units to make the produced gestures a continuous flow of movement. We empirically validated the use of gesture units in the generation and show that it causes the virtual human to be perceived as more natural.

**Key words:** Embodied Conversational Agents, Nonverbal Behavior Generation, Gesture Synthesis

## 1 Introduction

Researchers and users agree that interactive Virtual Characters (VCs) still lack naturalness in their movements, be it facial expression, gesture or posture changes. A recent state-of-the-art report [1] describes VCs as "wooden" and lacking variety in response, whereas human beings are all unique and have limitless variety. Although photorealism of computer graphics is steadily advancing "behavior may be more important than the visual realism of the character". Psychologists [2] and animators [3] alike confirm the important role of a visible and consistent personality for an animated character to make it more human-like and appealing. While research in the area of gesture synthesis has been active and successful for over two decades [4, 5] we believe that current research systems can be pushed in a variety of ways. First, most systems use a limited range of gestures and only a few systems can produce variants consistent with an individual style. Second, the produced gestures are rarely connected to form a fluid stream of gestures [6]. Finally, it is still hard to evaluate how natural the generated gestures are and to what aspects this naturalness is owed [7, 2].

In this paper we present a data-driven approach to synthesize gestures for a VC using procedural animation[5] [8–10]. Our approach features the recreation

[5] Video samples of animations can be found on http://www.dfki.de/~kipp/iva07

of the gesture style of a human performer, the production of a broad range of gestures, automatic synchronization with speech and the use of dynamic animation. While all technical details of the system are covered by Neff et al. [8], this paper can be regarded as a complement, giving a high-level overview, to then zoom in on gesture unit production. This sets the background for a new user study that validates the hypothesis that using gesture units makes the speaker appear more natural.

A *gesture unit* (g-unit) is a sequence of contiguous gestures where the hands only return to a rest pose at the end of the last gesture [11]. The concept stems from a proposed hierarchy of levels: on the lowest level, movement is seen as consisting of so-called *g-phases* (preparation, stroke, hold etc.), on the middle level these g-phases form whole gestures, called *g-phrases*, and on the top level gestures are grouped to *g-units* [12, 6]. In his extensive gesture research, McNeill [12] found that most of his subjects performed only a single gesture per g-unit most of the time; we call such gestures *singletons*. When we analyzed our speakers Jay Leno (JL) and Marcel Reich-Ranicki (MR), both well-known TV talk show hosts with active and appealing gesture behaviour, we found a different distribution (column $N$ displays the percentage of g-units containing $N$ gestures, first row taken from [12]):

| | **1** | **2** | **3** | **4** | **5** | **6** | **>6** |
|---|---|---|---|---|---|---|---|
| **McNeill's subjects** | 56 | 14 | 8 | 8 | 4 | 2 | 8 |
| **Speaker JL** | 35.7 | 15.7 | 17.1 | 5.7 | 11.4 | 5.7 | 8.6 |
| **Speaker MR** | 33.3 | 16.7 | 11.1 | 14.8 | 9.3 | 3.7 | 11.1 |

JL and MR preferred "longer" g-units, and we wondered whether this was one of the reasons why their gestures were much more interesting to watch than those of the average layperson. Therefore, we not only integrated the production of g-units in our gesture generation algorithm but also conducted a user study to examine the effect of g-units on the perception of the VC.

## 2 Related Work

Data-driven approaches based on motion capture can produce high quality movement, but motion variation is limited. Stone et al. [13] achieve some variability and gesture-speech synchrony by splicing and warping motion captured pieces, but as the authors indicate, the system does not have the generality of procedural animation. Our approach is also data-driven, but we annotate a higher level gesture representation that allows us to make use of the flexibility of procedural animation, rather than relying on replaying low-level motion data.

Kopp et al. [14] based their system on the *Sketch Model* [15] and can create gestures from arbitrary form specifications and handle co-articulation. While they focus on iconic gestures in spatial domains where the form-meaning relationship between speech and gesture is quite clear, we focus on metaphoric

gestures where this relationship is less clear and therefore, lends itself to statistical modeling. We share with the *Sketch Model* the processing of underspecified gesture frames that are gradually enriched in the planning process.

While many systems require a complex input specification, the BEAT system [16] works on plain input text and reconstructs linguistic features (e.g., theme/rheme) to generate synchronized gestures. Our system shares this general concept but instead of using hand-made, hard-coded rules for gesture generation we automatically extract them from video data of human speakers. Lee and Marsella [17] also systematically examine a video corpus but their generation rules are still hand-made. While they do not explicitly model personal style they take the *affective state* into account. However, compared with these two approaches, our animation engine allows a more fine-grained specification of gesture phase structure, similar to [18, 14]. Like our system, Hartman et al. [18] can generate *multiple strokes* as part of their expressivity parameters, but our system includes more detail concerning phase structure and timing. Several previous systems have been designed to animate expressive arm gestures (e.g. [19–21]). Our controller-based approach to physically simulated animation builds on similar work with hand-tuned controllers [22, 23] and using controllers to track motion capture data [24], but is the first animation system to use controllers to synchronize body movement with speech.

There is limited other work on modeling the gesturing style of a specific performer, although more abstract models are starting to appear [25]. We propose modeling specific human performers as a way to achieve more natural and varied behaviour. While sharing aspects with other gesture systems, we are the first to explicitly produce g-units and to evaluate the impact of g-units in a separate user study.

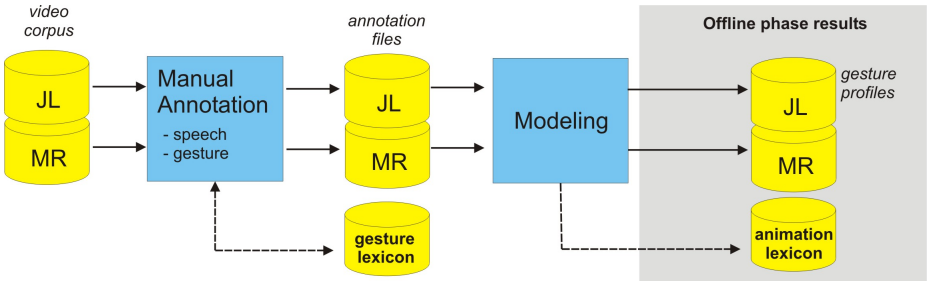


**Fig. 1.** Offline phase workflow: For every human speaker this work pipeline has to be completed once. The resulting data is used in the online system to generate gestures in the style of the modeled speaker for arbitrary input texts.

## 3 Gesture Generation and Animation

Our approach is mainly data-driven, using a video corpus of the human performer, but also incorporates general, character-independent mechanisms. A labour-intensive *offline* or preprocessing phase has to be completed once for each new human speaker (Figure 1). It results in a *gesture profile* of this particular speaker and an updated *animation lexicon*. The actual runtime or *online* system can then produce gestures in the style of one of the modeled performers for any input text (Figure 2).

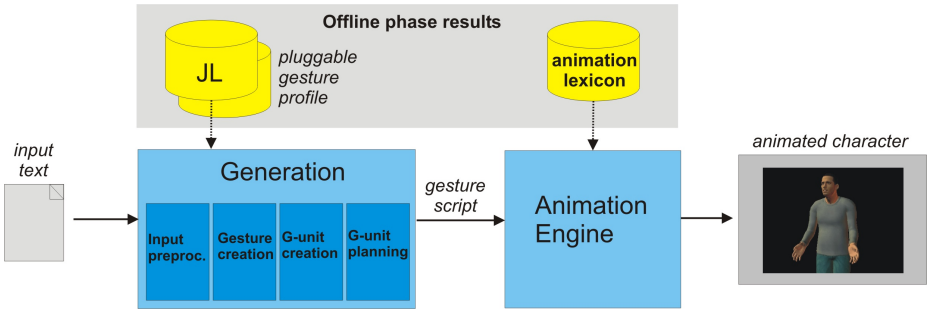


**Fig. 2.** Online processing: The runtime system can take arbitrary input texts (containing additional mark-up, see Section 3.3) and produce synchronized conversational gestures in the style of a modeled human performer. Gesture profiles can easily be exchanged and even mixed.

### 3.1 Offline: Annotation of Speech and Gesture

Our corpus consists of about 18 minutes of digitized video, 9 mins per speaker, from regular TV material where the speaker's face, torso and hands are visible. To prepare the video corpus for automatic analysis, a human coder transcribes speech and gestures according to our annotation scheme [9]. The transcription of this corpus results in 229 gestures (70 g-units) for speaker JL, and 192 gestures (54 g-units) for MR.

Coding starts with transcribing speech in the PRAAT[6] tool [27]. The transcript is imported to the ANVIL[7] tool [28] where gesture annotation is performed on three separate tracks, as shown in Fig. 4. On the first track, *g-phases* (preparation, stroke, hold etc.) are transcribed [6, 29]. In the track below the coder several of these phases into a *g-phrase* which basically corresponds to the notion of a "gesture" in everyday language.

On this track, the bulk of the annotation work takes place: selecting the lexeme, specifying form parameters and the link to speech. The *lexeme* refers to

[6] http://www.praat.org
[7] http://www.dfki.de/~kipp/anvil

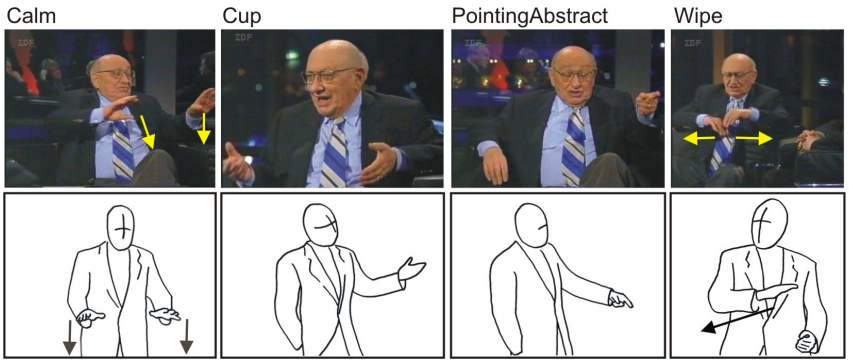


**Fig. 3.** Screenshots showing four lexemes from our gesture lexicon. The upper row shows samples performed by MR, the lower row shows schematic drawings (for readability) of speaker JL performing the same lexemes. The gesture lexicon contains additional textual descriptions on handshape, palm orientation, trajectory etc.

an entry in a *gesture lexicon* [10, 26] that we assembled beforehand, where 39 gestural prototypes, i.e. recurring gesture patterns, are described by form constraints and illustrations (Figure 3). These lexemes were partly taken from the gesture literature and partly extracted from our corpus. Of the 39 entries in our lexicon, 27 are used by both speakers, which is a large overlap that demonstrates a certain generality of the approach.

To describe gesture form the coder specifies handedness (RH, LH, 2H), trajectory (straight or curved), and the hand/arm configuration at the beginning and end of the stroke (only at the beginning of the hold for stroke-less gestures). The latter is specified with 4 attributes: three for the position of the hand and one for arm swivel. Finally, to mark the gesture's relation to speech we encode the *lexical affiliate* which is the word(s) whose meaning is most closely related to the gesture [30, 29] as a symbolic link to the respective word(s) on the speech track. For temporal reasoning we also encode the word(s) that co-occur with the gesture stroke. In the third track the gesture phrases are again combined to make up *g-units* [11, 6, 12]. See [9] for a full description of the coding scheme.

### 3.2 Offline: Modeling Gesture Behaviour

In the modeling phase we build a *gesture profile* from the annotations. First, the speech transcription is preprocessed to abstract away from surface text. The words are reduced to their word stem and then mapped to so-called *semantic tags* like AGREEMENT ("yes"), PROCESS ("create", "manage"), QUEST_PART ("why"), and PERS_NAME ("Michael Jackson"). The tags form a thin semantic

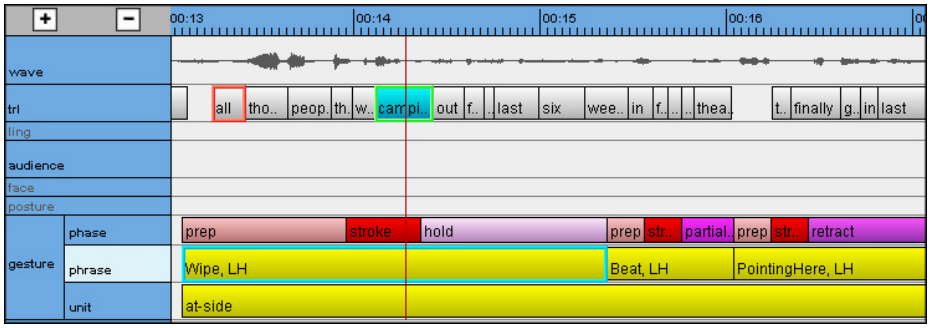


**Fig. 4.** Manual annotation was done in the ANVIL video annotation tool [28] on separate tracks for speech transcription and gesture annotation. Gestures were annotated on three hierarchically organized tracks for *phase, phrase* and *unit*.

layer between surface text and gestures, containing aspects of communicative function (e.g. AGREEMENT, QUEST_PART) and semantics (e.g. PROCESS).

After preprocessing we use the annotated *lexical affiliate* links between gesture and speech to compute the conditional probability that gesture $g$ occurs given semantic tag $s$. We also build a bigram model of gesture sequence, i.e. conditional probabilities that gesture $g_i$ follows $g_{i-1}$. We do the same for *handedness*, i.e. we model the probability that gesture $g$ is performed with the right, left or both hand(s), and store a bigram model of handedness sequences. Finally, we store the average number of multiple strokes and the average distance between stroke and co-occurring word per lexeme, and the general gesture rate (details in [8]). In the *GestureDB* we store every gesture occurrence, together with the spatial data from the manual annotation. It is *not* a collection of animation clips but rather contains essential form data about the gesture's expressive phase. This data includes spatial "pass through" points, indicating hand positions at the beginning and end of the stroke, and the shape of the movement trajectory (straight or curved).

In conjunction with the gesture lexicon we created an *animation lexicon* where animation-relevant data is stored for each lexeme, including palm orientation and posture changes, as well as warps to the motion envelope. For each new speaker, the lexicon can be (manually) enriched with character specifc information, including a default posture and succession pattern. The animation lexicon also defines the form of the *after-strokes* in a multiple stroke. A *multiple stroke* in a gesture contains many strokes: We call the first of these strokes *main stroke* and the subsequent ones *after-strokes*. After-stroke information generally involves small hand and forearm movements and specifying whether a hold phase is present.

### 3.3 Online: Generation of Gestures and G-Units

The online system transforms novel text into an animated sequence of gestures. The input text must be marked-up with temporal word boundaries, utterance

segmentation and annotation of rheme and focus [31, 16]. Currently, we add this information manually. The text is transformed to a graph where nodes represent time points and arcs represent words or gestures. Words are stemmed and mapped to a semantic tag, just as in the modeling step. The generation then proceeds in two steps: gesture creation/selection, and g-unit formation.

**Gesture creation and selection** Using the speaker's gesture profile, we first create a large number of underspecified gesture frames which are then thinned out by a selection criterion. Gesture candidates are all gestures whose conditional probability to co-occur with a semantic tag exceeds a threshold of 0.1. One copy of the gesture is placed over the semantic tag in the graph. If the semantic tag is within a rheme, another copy is placed on the *focus* of the rheme [31]. For selection, we utilize the *n-gram* models to construct a sequence with maximum likelihood that observes the speaker's gesture rate. We use a similar approach to determine the handedness. Handshape is added according to our animation lexicon. After this step, we have a sequence of underspecified gesture frames that contain lexeme, handedness and handshape. The position in the graph determines the word(s) the gesture has to be synchronized with.

**G-unit formation** The gesture hierarchy of phases, phrases and units is hypothesized to correspond to levels of speech phrase organisation. For instance, Kendon [11, 32] suggested a correlation between *intonation units* and g-units. Such concepts go back to the hypothesis that speech and gesture originate from a single source, called *growth point* [12] or *idea unit* [32]. In our algorithm we try to approximate these concepts.

We "grow" g-units by merging neighbouring gestures (distance $< 1.5$ sec) within a single utterance segment, taking utterance segmentation as a crude approximation of intonation units. Note that the distance criterion is not speaker-dependent but could be made so to model the fact that speakers prefer longer or shorter g-units. Now, gesture phases, spatial parameters, and gesture-speech timing remains to be specified. Phase structure is determined as follows: If, for two consecutive gestures $g_{i-1}$ and $g_i$, there is time for a preparation (.5 sec), then insert one. If not, insert a spatial constraint marker that $g_i$'s start position must match $g_{i-1}$'s end position. Now find a suitable gesture, using lexeme and spatial constraints, from GestureDB (randomize over possible options) and generate multiple strokes randomly using the speakers mean value and standard deviation for this lexeme. Resulting temporal conflicts with the succeeding gesture $g_i$ are resolved by either moving $g_i$ back in time (up to a certain limit) or eliminating it altogether.

We synchronize gesture and speech by positioning the end of the stroke at the end of the corresponding word, using a random offset based on the speaker's mean value. For multiple strokes we synchronize all after-strokes with word end times of subsequent words, enforcing a minimum time span. For stroke-less gestures we synchronize the *start* of the independent hold [6] with the start of the word. In a final wrap-up phase we subtract 0.3 sec from all main stroke times

(0.12 sec for after-strokes in a multiple stroke) as we empirically found this offset necessary to make the gesture timing look natural, and fill in all gaps within a g-unit with holds.

As Kendon [11] pointed out, the retraction phase is a property of the g-unit and *not* of a single gesture because within-unit gestures cannot have a full retraction by definition. We therefore generate a retraction phase for each g-unit using a simple heuristic to determine the rest position: if the distance to the following g-unit is small, retract to 'clasped', if medium retract to 'at side', if large retract to 'in pockets'. The results of the gesture generation are written to a *gesture script* containing the lexeme, timing and spatial information computed above.

### 3.4 Procedural Animation

The role of the animation engine is to take the gesture script as input and output a final animation. It does this in three main steps. First, it computes additional timing data and adds information from the animation lexicon (Section 3.2). Second, it maps all data to a representation suitable for generating animation. This representation is essentially a keyframe system and has tracks for every Degree of Freedom (DOF) of the character's body. The tracks are populated with desired angles at specific points in time and transition curves that specify how the DOF values should change over time. The basic keyframe representation is augmented with offset tracks (see below) that are summed with the main tracks to determine the final desired values for the joints. Additional tracks are used to control real time behaviour, such as gaze tracking. Once this representation has been populated, the third and final step is to generate the actual animation. The system can generate either dynamic, physically simulated skeleton motion, or kinematic motion. The system also produces eye brow raises on stressed phonemes and lip synching . Some of the key components of the system will be described below. Further details can be found in [8].

The gesture script does not contain all the data necessary for defining an animation and is hence augmented by the animation engine. The first step in this process is to complete the timing information. The animation engine also resolves possible spatial conflicts. These occur due to the coarse-grained spatial annotation, which can fail to record small hand separations and movements. Offsets are automatically added to deal with these cases. The rest of the augmentation involves adding the gesture-specific data from the animation lexicon, all of which is taken into account when the gesture poses are solved for and written into the low-level representation.

A pose is calculated at each phase boundary (e.g. the start and end of a stroke) using a series of local IK routines for the arms, wrists and lower-body chain. This is augmented with a feedback based balance controller and automatic collarbone adjustment based on the height of the hands. The wrist targets are defined in a body relative way, which is important for allowing the gestures to move appropriately with the body. Realtime processes provide gaze tracking and balance adjustment.

The offset tracks are used to layer different specifications together, such as when combining body rotation with gestures. They are also used to curve the spatial trajectory of motions by adding offsets in joint space. This allows a stroke to be specified with a straight or curved trajectory. It also allows the creation of more complicated gestures such as progressives, which feature the forearm and hand making circular motions.

When computing physically simulated animation, we use a controller based approach whereby an actuator is placed at each DOF which calculates the torque required to move the character towards the desired configuration. We use an antagonistic formulation of proportional-derivative control, following [33]. In order to synchronize the gesture timing with the specified timing and to preserve the spatial extent of the motions, we both shift the start time earlier and shorten the duration of each movement when doing dynamic simulation. The dynamic animations contain additional effects due to momentum, such as pendular motion when a character brings his hand to his side or wrist overshoot when the forearm is moved rapidly. A variable time step Rosenbrock integrator [34] is used to compute the motion using simulation code from SD/Fast [35]. Whereas kinematic animation runs in real time, physical simulation is only about one tenth real time so must be pre-computed and played back when needed.

### 3.5 Validation: Recognizability of Gesture Style

We validated the system with an emprical study where 26 subjects were exposed to video clips of generated gestures [8]. After a short training phase where the subjects watched original clips of the performers JL and MR, they went through two tests. In Test 1 the subjects watched a single animation, produced on a novel, synthesized text (prologue from "Star Wars"), and were asked which performer (JL or MR) it was based on. In Test 2 they saw a side-by-side screening of animations based on each model and were asked to identify which was which. In Test 1, subject selected the correct performer 69% of the time ($t(25)=2.083$; $p < .05$). In Test 2, subjects achieved, as expected, a higher success rate of 88% ($t(25)=6.019$; $p < .001$). These results show that the subjects were able to recognize the human original, based on only the gesture behaviour.

## 4 Evaluating Gesture Units

Despite the encouraging validation of our approach, evaluating nonverbal behaviour of VCs remains a difficult task [7, 2]. One strategy is to single out one parameter in the generation process. In our approach, a major decision was to generate g-units instead of singletons. So our underlying hypothesis was: Using g-units makes our VC look more natural.

To test the impact of g-units beyond naturalness we selected a number of dimensions for personlity perception. We picked three dimensions from the "Big Five" [36] (friendliness, nervousness, extrovertedness) and added dimensions of potential relevance for VC applications, and arrived at 6 dimensions: naturalness, friendliness, nervousness, extrovertedness, competence, and trustworthiness.

### 4.1 Method

25 subjects (12 female) participated in this experiment. 14 subjects were recruited in Germany, 11 in the US.

**Material** We prepared video material with a virtual character gesturing in two different ways for the same verbal material[8] (some English, some German). In the **U**(nit) version we used the g-units generated by our system. In the **S**(ingleton) version we modified the animations so that only singleton gestures occur. For this, we replaced every hold that separated two gestures with a retraction. We left multiple strokes intact where possible and only removed enough after-strokes to make room for a retraction. In both versions, we replaced all rest poses with the "hands at side" rest pose, to make the effect of returning to rest pose clearer. We cut the material to 11 pieces of length 11-28 sec for each version (total of 2:58 min). In each piece we tried to include 2-3 g-units since in pilot studies too short clips did not show the effect of g-units clearly enough, as frequently returning to rest pose does not seem odd in a short clip. However, much longer clips would have made it hard to remember the impression of the first clip after viewing the second. We considered a side-by-side presentation which makes the difference between the two versions very clear but weakens the generality of the result. Pilot studies also caused us to block out the face of the VC (see Figure 5) since subjects reported being quite absorbed by facial expressions and head movement (gaze). The 11 clip pairs were presented in random order. Within-pair order (S-U or U-S) was balanced across subjects.


**Fig. 5.** In the g-unit experiment the face was blocked out to reduce distraction.

**Procedure** In the instruction the subject was informed that s/he was to compare two versions of the same virtual character and decide in which of the versions the character's behaviour was more natural, more friendly, more nervous, more extroverted, more competent, and more trustworthy. Subjects were explicitly requested to judge "intuitively" and to focus on the bodily behavior.

Each subject went through 11 rounds. In each round, both animation versions, labeled "A" and "B", were shown consecutively. Afterwards, a questionaire

[8] Samples of this material can be found on http://www.dfki.de/~kipp/iva07

appeared with the options "A", "B" or "undecided" for each dimension. How much of the speech was understood was asked on a 3-point scale. In each round, the two versions could be played repeatedly. After the experiment the subjects were interviewed about perceived differences and answering strategies.

### 4.2 Results

Since our subjects came from two countries (US/Germany) we checked for differences between the two populations with a Mann-Whitney U-test which did not reveal differences on any of the dimensions. Separating the data according to "speech understanding" did not reveal any differences (2-factorial ANOVA).

Looking at all clips and all subjects[9], the **U** version was selected significantly more often (60% vs. 40%) to be *more natural* than the **S** version (1-tailed t-test[10]: $t(24)=3.1062$; $p < .01$). For the other five dimensions we found the following results: Subjects picked the **U** version significantly more often as *more friendly* (2-tailed t-test: $t(24)=4.2774$; $p < .001$), *more trustworthy* ($t(24)=4.3085$; $p < .001$), and tendentially more often as *more competent* ($t(24)=1.7220$; $p = .10$). The **S** version was perceived significantly more often as *more nervous* ($t(23)= 3.7999$; $p < .001$)[11]. There was no significance either way for *extroversion*.

We then tested whether any of the dimensions were correlated in the sense that subjects consistently made the same decision on two dimensions (e.g., natural and friendly). We conducted a unidimensional $\chi^2$ test[12] on all possible dimension pairs. The results (Table 1) show that naturalness, friendliness, competentence and trustworthiness are positivly correlated with each other, whereas nervousness is negatively correlated with all of these. Extrovertedness is only directly correlated with nervousness.

| | **friendly** | **nervous** | **extrovert** | **competent** | **trustworthy** |
|---|---|---|---|---|---|
| **natural** | (+) .89 | (−) .20 | .49 | (+) .83 | (+) .92 |
| **friendly** | | (−) .17 | .44 | (+) .75 | (+) .85 |
| **nervous** | | | (+) .62 | (−) .23 | (−) .19 |
| **extrovert** | | | | .54 | .48 |
| **competent** | | | | | (+) .83 |

**Table 1.** Relative frequencies of equal decisions when comparing dimensions. (+) means that subjects made the same decision, (−) means they made the opposite decision *significantly* often (all highly significant, $p < .001$; $\chi^2$ values omitted for readability).

### 4.3 Discussion

Using gesture units makes a virtual character look more natural. Our results clearly confirmed this hypothesis. However, our interviews revealed that the

[9] The "undecided" category was not considered in the analysis. For dimension naturalness, we had 14% "undecided" cases. For all other dimensions 20–28%.

[10] Since the data was normally distributed (verified with Kolmogorov-Smirnov one-sample test/Lilliefors probabilities), parametric methods were applicable.

[11] One subject chose "undecided" on this dimension for all clips.

[12] Because of the small sample we could not assume normal distribution of the basic population.

effect is very subtle. Only one of the subjects was able to explicitly tell the difference between versions. Most subjects said that the difference was very hard to see, some plainly said they saw no difference. Those who saw differences assumed that they were in the timing, extent or smoothness of movements, synchronization with speech or torso deformations – these aspects were exactly equal in both versions.

Our collected evidence suggests that for creating a friendly, trustworthy and natural VC, the use of g-units plays a subtle yet important role. G-units may also have a positive impact in terms of competence. If a more nervous or even unnatural character is desired, singleton gestures should be preferred. However, our experiment only discriminates between singleton gestures and g-units of length $> 1$. More precise effects of g-unit length remain to be examined. It is interesting to note that in the context of "gestural behaviour of virtual characters" the dimensions naturalness, friendliness, trustworthiness and competence seem to form one cluster where one implies the other. Nervousness stands in an inverted relationship with each of these, and finally, extroversion is a dimension that was left totally unaffected by the g-unit condition.

Although our goal was to make *virtual* humans look more natural, our results may have implications for *real* humans. For instance, based on our results, a rhetorics teacher could recommmend her students to "connect" gestures in order to appear less nervous. While this needs to be complemented by studies on how far the results with virtual characters transfer to human-human interaction [2], a VC may prove an ideal research tool for the social sciences [37–39].

## 5 Conclusion

We presented a data-driven approach to gesture synthesis that allows the synthesis of gestures on novel input text in the style of a modeled human performer, including a validation study on the "recognizability" of the produced gesture style. The approach not only creates a broad range of gestures but also connects the generated gestures into *gesture units* to produce a smooth and fluid stream of gestures.

We validated the positive effect of producing g-units as opposed to producing mere singleton gestures in a user study and found that the g-unit version was perceived as more natural, more friendly, more trustworthy and less nervous. While the study confirmed our hypothesis that generating g-units is a definite advantage, there is much room for further empirical exploration. The effect of different g-unit lengths and also decisions on handshape selection or the use of multiple strokes and various animation parameters could be studied to arrive at a more complete set of validated principles for gesture synthesis.

**Acknowledgments.** We would like to thank all the subjects who participated in our user studies, and Nuance Communications Inc. for providing us with a text-to-speech synthesis software.

## References


1. Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., Slater, M.: Building Expression into Virtual Characters. In: Eurographics Conference State of the Art Report, Vienna (2006)
2. Nass, C., Isbister, K., Lee, E.J.: Truth is beauty: Researching embodied conversational agents. In Cassell, J., Sullivan, J., Prevost, S., Churchill, E., eds.: Embodied Conversational Agents. MIT Press, Cambridge, MA (2000) 374–402
3. Thomas, F., Johnston, O.: The Illusion of Life: Disney Animation. Hyperion Press, New York (1981)
4. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated Conversation: Rule-Based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents. In: Proceedings of SIGGRAPH '94. (1994) 413–420
5. Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., Petajan, E.: Creating interactive virtual humans: Some assembly required. IEEE Intelligent Systems (July/August 2002) 54–63
6. Kita, S., van Gijn, I., van der Hulst, H.: Movement phases in signs and co-speech gestures, and their transcription by human coders. In Wachsmuth, I., Fröhlich, M., eds.: Gesture and Sign Language in Human-Computer Interaction, Berlin, Springer (1998) 23–35
7. Ruttkay, Z., Pelachaud, C., eds.: From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer (2004)
8. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style. Transactions on Graphics (2007) accepted.
9. Kipp, M., Neff, M., Albrecht, I.: An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. Journal on Language Resources and Evaluation - Special Issue on Multimodal Corpora (2007)
10. Kipp, M.: Gesture Generation by Imitation: From Human Behavior to Computer Character Animation. Dissertation.com, Boca Raton, Florida (2004)
11. Kendon, A.: Gesture – Visible Action as Utterance. Cambridge University Press, Cambridge (2004)
12. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)
13. 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: Proc. SIGGRAPH 2004. (2004) 506–513
14. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds **15**(1) (2004) 39–52
15. de Ruiter, J.P.: The production of gesture and speech. In McNeill, D., ed.: Language and Gesture: Window into Thought and Action. Cambridge University Press, Cambridge (2000) 284–311
16. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: Proceedings of SIGGRAPH 2001. (2001) 477–486
17. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Proc. of the 6th International Conference on Intelligent Virtual Agents, Springer (2006) 243–255
18. Hartmann, B., Mancini, M., Buisine, S., Pelachaud, C.: Design and evaluation of expressive gesture synthesis for ecas. In: Proc. AAMAS. (2005)

19. Chi, D.M., Costa, M., Zhao, L., Badler, N.I.: The EMOTE model for effort and shape. In: Proc. SIGGRAPH 2000. (2000) 173–182
20. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. In: Gesture Workshop 2005. (2005)
21. Neff, M., Fiume, E.: AER: Aesthetic Exploration and Refinement for expressive character animation. In: Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2005. (2005) 161–170
22. Hodgins, J.K., Wooten, W.L., Brogan, D.C., O'Brien, J.F.: Animating human athletics. In: Proc. SIGGRAPH 1995. (1995) 71–78
23. Faloutsos, P., van de Panne, M., Terzopoulos, D.: The virtual stuntman: Dynamic characters with a repertoire of autonomous motor skills. Computers & Graphics **25**(6) (2001) 933–953
24. Zordan, V.B., Hodgins, J.K.: Motion capture-driven simulations that hit and react. In: Proc. ACM SIGGRAPH Symposium on Computer Animation. (2002) 89–96
25. Noot, H., Ruttkay, Z.: Gesture in style. In: Proc. Gesture Workshop 2003. Volume 2915 of LNAI., Berlin, Springer (2004) 324–337
26. Webb, R.: Linguistic Properties of Metaphoric Gestures. UMI, New York (1997)
27. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (version 4.3.14) [computer program]. Retrieved from http://www.praat.org/ (2005)
28. Kipp, M.: Anvil – a Generic Annotation Tool for Multimodal Dialogue. In: Proceedings of Eurospeech. (2001) 1367–1370
29. McNeill, D.: Gesture and Thought. University of Chicago Press, Chicago (2005)
30. Schegloff, E.A.: On some gestures' relation to talk. In Atkinson, J.M., Heritage, J., eds.: Structures of Social Action. Cambridge University Press (1984) 266–296
31. Steedman, M.: Information structure and the syntax-phonology interface. Linguistic Inquiry **34** (2000) 649–689
32. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. In Key, M.R., ed.: Nonverbal Communication and Language. Mouton, The Hague (1980) 207–227
33. Neff, M., Fiume, E.: Modeling tension and relaxation for computer animation. In: Proc. ACM SIGGRAPH Symposium on Computer Animation 2002. (2002) 81–88
34. Press, W.H., Tukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: The Art of Scientific Computing. 2nd edn. Cambridge University Press (1992)
35. Hollars, M.G., Rosenthal, D.E., Sherman, M.A.: SD/FAST User's Manual. Symbolic Dynamics Inc. (1994)
36. McCrae, R.R., John, O.P.: An Introduction to the Five-Factor Model and Its Applications. Journal of Personality **60** (1992) 175–215
37. Martin, J.C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C.: Multimodal Complex Emotions: Gesture Expressivity And Blended Facial Expressions. Special issue of the Journal of Humanoid Robotics (3) (September 2006) 269–291
38. Krämer, N.C., Tietz, B., Bente, G.: Effects of embodied interface agents and their gestural activity. In: Proc. of the 4th International Conference on Intelligent Virtual Agents, Springer (2003)
39. Frey, S.: Die Macht des Bildes: der Einfluß der nonverbalen Kommunikation auf Kultur und Politik. Verlag Hans Huber, Bern (1999)