Towards Higher Quality Character Performance in Previz

Stacy Marsella* Ari Shapiro† Andrew Feng‡ Yuyu Xu§ Margaux Lhommet¶
Stefan Scherer‖
Institute for Creative Technologies
University of Southern California

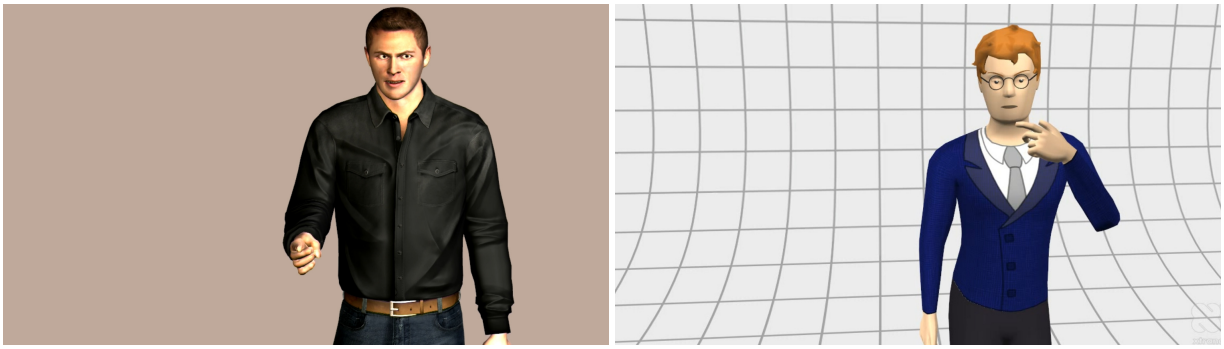

Figure 1: *Comparison of previsualization results generated from our system (Left) and Xtranormal (Right). Our method produces semantically appropriate behavior by performing deep analysis of syntactic, semantic and rhetorical structures of the utterance.*

Keywords: previsualization, animation, gestures

1 Motivation

Previsualization tools are used to obtain a preliminary but rough version of a film, television or other production. Used for both live-action and animated films, they allow a director to set up camera angles, arrange scenes, dialogue, and other scene elements without the expense of paying live actors, constructing physical sets, or other related production costs. By seeing an early approximation of the final production, decisions about scenes, elements, story and the factors affecting it can be made early in the process, potentially reducing costs and improving overall quality. Current previsualization technologies have made inroads into generating these "videomatics", where controls over cameras and static elements, such as buildings, roads and scenery, can be quickly incorporated from a low cost libraries of 3D assets. Even the generation of effects such as explosions, running water, and smoke can be quickly generated in previz scenes from commodity software.

However, the generation of a character's body and facial performance can be difficult to create quickly without the assistance of animation experts to create reasonable postures, gaze lines and positioning that reasonably reflect the desired performance in a scene. Tools for automated performance generation can be used to record and playback vocal performances by performing automated lip synchronization to speech along with limited facial expression. However, these synthesized animated performances don't reflect the content of the utterance, and typically lack the emotion related to the recorded voice performance. Thus, costly and time consuming manual intervention from previsualization experts is still required to generate a character performance from audio. Indeed, generating any convincing performance of a digital character is an expensive, time consuming task that requires expert intervention.

In this work, we seek to raise the standard of quality for automatic and minimal cost 3D content by addressing a difficult problem; the automated generation of a 3D character performance from audio. Our goal is to minimize the role of expert assistance in the generation of 3D previsualization of characters by providing a high quality character performance that is generated from an audio recording. Numerous technologies with various cost-to-quality issues could be used to create medium and high quality character performances. For example, the addition of a performance capture system for the face and/or body could be used to create preliminary versions useful for previsualization. However, this would represent an addition of hardware, software, as well as performance capture software expertise and actors, which in turn increases cost. By contrast, we are seeking an increase in quality without a proportionate increase in cost. One of the most costly factors is the need for an animation expert to construct a believable performance with a previz tool. Thus we seek to automate character animation with minimal input and without expert animation assistance.

2 Contribution

Our method synthesizes an animated performance for a 3D character automatically. The user only needs to provide input audio and a transcription of the spoken text as input. The resulting performance is a combination of gestures, eye gaze, head movements, facial expressions, and lip-syncing animations.

*e-mail:marsella@ict.usc.edu
†e-mail:shapiro@ict.usc.edu
‡e-mail:feng@ict.usc.edu
§e-mail:yxu@ict.usc.edu
¶e-mail:lhommet@ict.usc.edu
‖e-mail:scherer@ict.usc.edu

Our contribution is the addition of deep semantic and rhetorical analysis of the spoken utterance, which allows the generation of nonverbal behaviors that are appropriate for the utterance. In contrast, other methods rely on prosody (stress or intonation of the speech) or use simple rules that trigger gestures based on word spotting [Cassell et al. 2001; Niewiadomski et al. 2009]. Work by [Neff et al. 2008] requires additional annotation for the new input audio and text. Moreover, since both the input audio and text are analyzed, the resulting gesture performance will not only match the timing of words, but will also match the context of the sentence. This helps our system generate more appropriate gesturing results than methods that based on prosody[Levine et al. 2010; Levine et al. 2009] or key words [Bergmann and Kopp 2009].

We demonstrate in our video the effectiveness of such analysis. In addition, we submit a study that shows that our method has superior performance to those that use prosody alone, or that use random gesturing.

3 Method Overview

Figure 2 gives an overview of our system.

3.1 Audio Processing

Given an input audio and its corresponding text, our system first analyzes the audio signal to determine the emotional state for the overall utterance as well as which words are being stressed. The emotional state is determined from the agitation level [Scherer et al. 2013] which ranges from tense to lax, and then categorized as low, medium or high. The audio signal is then analyzed to generate a forced alignment of phonemes according to the text transcription. This phoneme schedule is then used to drive an automated lip-syncing based on a diphone-driven method [Xu et al. 2013]. The acoustic signal is then analyzed to locate audio segments with high fundamental frequency, which are then mapped to the word location, and assign a stress score for each word spoken.

3.2 Syntactic, Semantic and Rhetorical Analysis

The text associated with the sentence is then goes through a ranges of analyses beginning with a parse to derive its syntactic structure. Analyzing the syntactic structure overcomes the limitation from other methods that use simple, single word identification, allowing more sophisticated analysis of the text at the level of phrase and clause structures and allowing the synchronization of behaviors to span the entire phrase, rather than individual words. For example, a sweeping gesture and/or head motion can be associated with the entire quantified noun phrase, "every person on the team", instead of simply the quantifier "every".

The system then determines the communicative functions realized by the utterance, by applying pattern matching to infer semantic, metaphoric and rhetorical elements relevant to non-verbal behavior. For example, the term "much more important" will be identified as *strong positive comparative*. The terms "very" or "extremely" are identified as intensifiers. Phrases such as "big idea" that suggest physical properties (either actual or metaphorically) are identified. Mental states are identified, including cognitive load (inferred from dysfluencies in the text) or emotion (based on audio processing). In all, we currently employ a dictionary of 170 key English words along with 91 rules associated with those words. The rules are then mapped to suitable non-verbal behaviors with proper timing associated with words and terms in the sentence. The rules are generated from expert knowledge in human nonverbal behavior, and include rules such as using beat gestures to emphasize comparatives, to saccades of the eyes upwards during a verbal dysfluency (such as a spoken 'umm' or 'err') to represent cognitive load. Once we have the desired behaviors and their corresponding timing, our animation system will process the scheduled behavior to synthesize the output animations. Behaviors are assigned priority values which dictate preference of one over the other. These priorities are influences by the audio processing that identified stressed words. Further details can be found in [Marsella et al. 2013].

One of the key elements in our system is to synthesize a comprehensive character performance including gestures, gazes, head movements and so forth [Shapiro 2011]. Here we procedurally generate the gaze and head movements. Specifically, gaze is generated by rotating spine, neck, and eye joints to move the line of sight toward the target. Head movements are generated by adding phase-shifted sine wave along each rotation axis on the neck joint.

Gesture synthesis is based on a set of gesture motions. Our system requires approximately 29 types of gestures for each of three emotional states. Each gesture type includes both the single handed and double handed gesture motions. To reduce the number of required gestures, we mirror the single handed gesture to produce the equivalent gestures for the other hand. Moreover, when the timing between two adjacent gestures are close to each other, it would be unnatural to simply playback both gesture separately. Instead, the system will automatically connect two gestures by introducing a *gesture hold* for the first gesture to allow a smooth transition to the next gesture. This increase the variations in the results and creates a more organic feel for the synthesized gesture animations.

We generate automatic listening behaviors for the other (nonspeaking) characters in the scene by following two rules: nodding in response to an utterance, and mimicking (mirroring) head movements. These two rules are sufficient to generate reasonable reactions from nonspeaking characters.

Figure 3: *Study comparing the same utterance using (left) our method, (middle) prosody-based beats, and (right) random gestures.*

We perform a study of 69 participants to show the importance of proper gestures in conveying the intentions of speaker. Here we compare the gesture performance synthesized from our system with the results based on only prosody, and the results based on randomly selected gestures. We produce the animations from all three methods and ask the study participants to rank the quality of each videos. Figure 3 shows an example of gesture variations between three methods. Based on the study results in Figure 4, most study participants (71%) prefer the results from our method comparing to the results based on prosody (19%) and randomly selected gestures (10%). This study shows that our method is capable of producing a closer approximation of a final video, and thus is more suitable for pre-visualization application than previous methods.

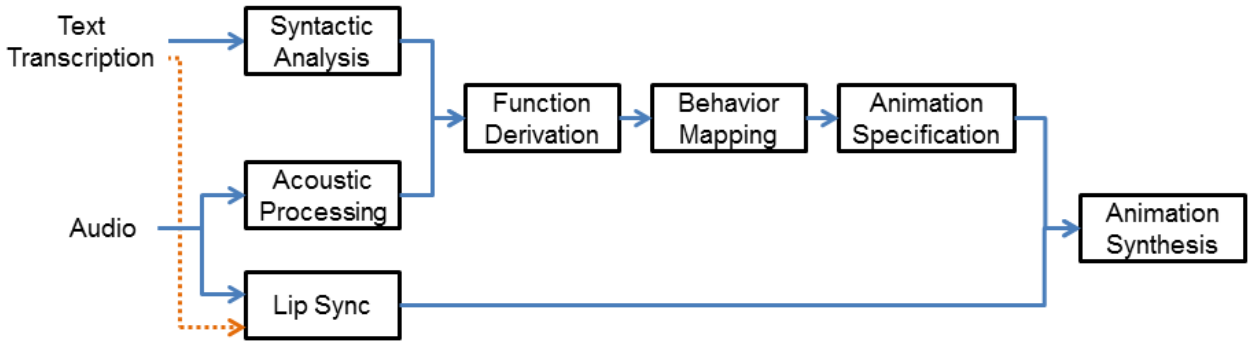

Figure 2: *Overview of the our method. Our system accepts an audio signal and a transcription of the audio as input and generates an animated performance as output.*

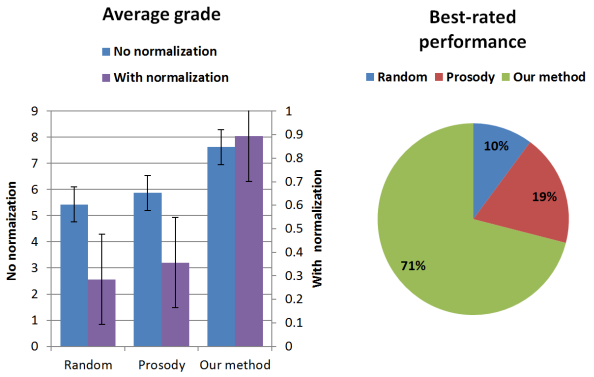

Figure 4: *Comparison of appropriateness between random gestures, prosody-based beats and our method's semantic gestures.*

4 Comparison with Existing Tools

To our knowledge, few tools exist that produce a convincing animated performance of a virtual character with minimal user inputs. Source Filmmaker [SFM 2013] is a machinema editing tool that can allow the user to create and edit the recorded gameplay footage. It is a full feature tool that allow fine user control in editing the animation result. However, since its purpose is for performance editing instead of synthesis, it is not suitable for producing the character performance in pre-visualization for emotionally charged conversations.

Xtranormal [Xtranormal 2013] is a tool that allows the user to quickly generate character performance from words in a script. It provides a simple interface that allows the users to position the characters, add the dialog, and adjust the camera shots. The character performance can then generated in a semi-automatic fashion. Their system will automatically produce the lip-syncing from the audio. It also produces gesturing and gaze animations based on simple conditions such as punctuation marks and the positions of nearby characters. The user is also allowed to manually adjust the performance by annotate the dialog text with specific gestures or gazes.

We compare our system with Xtranormal by producing the previsualization examples of a movie sequence from both systems. Both system are provided with the same level of inputs including the audio clips, texts, and the emotional state of the actor for each sentence. As shown in Figure 1 and the accompanied video, our system produces a performance that is more reminiscent of the final performance by comparison. Although this is partially due to the fact that we use higher quality assets, we also observe that the choice of gestures and other performances from Xtranormal is not appropriate to the context of the sentence. Since their system based on only simple syntactical rules (such as an exclamation point at the end of a sentence generates a beat gesture), it lacks a deep understanding of the meaning of the utterance. Thus, its results are not likely to well represent the utterance without additional user inputs.

5 Performance Editing

The goal this work is to automate as much of the performance as possible. However, a useful previz system must include the ability to edit or change the resulting animation without encountering low-level constructs such as animation curves or IK chains, and in turn requiring an expert animator or similarly skilled user.

Several automated procedures can be manually changed as needed. For example, the emotional state of the character is automatically determined from the tenseness of the audio, categorized as low, medium or high, and mapped to an equivalent emotional state. However, the user can override this value by manually setting the emotional state of the character directly. Word stresses are determined by the fundamental frequency of the audio, but can be marked with particular stress values, which would in turn trigger rules for stressing words. In addition, character relationships can be specified which affect listening behavior. For example, annotating one character as an antagonist of another can switch the listening behavior from nodding in response to an utterance, to shaking the head in response to an utterance. In addition, activation of head movements of one character in response to another can be suspended in this circumstance.

Our method automatically generates a description of a set of timed behaviors for a character. The individual behaviors are specified at a coarse level, called behavior blocks, which indicate the type of behavior (head movement, eye movement, gaze, gesture, facial expression, and so forth) along with specific timing and characteristics of such behaviors, such as gesture type, gaze direction, activation time, and so forth. This is in contrast to traditional animation tools which define movement as a set of curves, or curve controls.

The animation system then interprets these instructions and synthesizes an animation that satisfies the behaviors and timing constraints. By specifying the behavior instructions in a human-readable, high level format, the previz user can then modify and can then modify the character performance by adding, removing

Figure 5: *Previsualization results generated from our system. Our method produce an entire character's performance from audio signal.*

or changing behaviors to further customize them within an editing tool, seen in Figure 6. Interpretation of these behaviors includes rules to play, drop or coarticulate movements together based on their proximity. For example, when performing multiple gestures in close temporal proximity to each other, the gestures may be shortened or modified to accommodate the movement of the other gestures. Gestures in close proximity to each other are automatically modified to include periods of holding, and subsequent transitions to meet the constraints of the subsequent gesture.

This coarse definition of behaviors allows the previz designer to move, change or remove entire blocks of behaviors in order to modify the performance without needing knowledge of the underlying control mechanisms. Since the behavior blocks are translated into a set of animation curves, it would also be possible expose such animation curves for editing to an animator in order to achieve a specific performance as part of a refinement step.

6 Discussion

We have discovered that a completely automated animated performance requires an integration of several simultaneous behavior layers in order to be effective. Lip synching, emotional expression, head movements and gestures must coexist and coordinate with each other in order to generate a convincing performance. For example, automatically synthesizing an animation on a character simply asking the question 'Why?' require coordination of a raised eyebrow for questioning, lip syncing to the word, moving the head, while simultaneously raising an open hand. Also, subsequent behaviors must interact smoothly with existing ones, requiring intelligent decisions about which behaviors to retain, ignore or coarticulate.

Currently, our system utilizes three gesture sets for each character, representing low, medium and high energy, with 29 gestures in each set. These gesture sets are then retargeted to new characters. However, people gesture and move in different ways. Thus additional gesture sets are required to express different personalities and styles. It is conceivable that such styles could be generated from performance capture or through hand-construction. In addition, the behavior rules can be customized per character to achieve variation in interpretation.

A key to the system is the large amount of variation that can be achieved from a single utterance. Despite having identical word contents, two utterances can vary greatly in their vocal performance, which in turn produces variations in timing, stress and emotional content. Such changes can in turn trigger different behaviors, such as emphasized, energetic gestures over soft, gentle ones. Choosing different behaviors can in turn cause different choices in coarticulation by the animation system (such as dropping or extending the timing of a previous gesture) which, in turn, changes the animated performance.

A previsualization effort can span a range of approaches and resource needs [Hetherington 2006]. A high-end previsualization can represent all the elements needed for final production, including important character movements, timings, backgrounds, and so forth. At the other end of the spectrum, a low-end previsualization effort could include only camera blocking and still images. Thus our automated character animation solution may not be useful to some previsualization approaches that do not incorporate audio, or to those that require specific character motion that is unrelated to the content of the audio. Since our system infers the emotional state of the characters from the content of the utterance, it can fail to detect complicated mental processes that are unrelated to the utterance, such as sarcasm.

Clearly, superior animated performances can be generated with more effort, expertise and time. However, the goal of this work is to generate a reasonably high quality animated performance with little effort or animation expertise. We believe that by doing so we are able to show a better likeness of the final production, and thus increase the usefulness of the previsualization process. We consider this effort to be one of many layers needed in the previsualization process.

Acknowledgements

Thanks to Teresa Dey for her work generating the animation assets needed to synthesize the animated performances.

References

BERGMANN, K., AND KOPP, S. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 361–368.

CASSELL, J., VILHJLMSSON, H. H., AND BICKMORE, T. 2001. BEAT: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 477–486.

HETHERINGTON, J., 2006. The prevalence of previs, animation world network, Jan.

LEVINE, S., THEOBALT, C., AND KOLTUN, V. 2009. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph. 28*, 5 (Dec.), 172:1–172:10.

LEVINE, S., KRÄHENBÜHL, P., THRUN, S., AND KOLTUN, V. 2010. Gesture controllers. In *ACM SIGGRAPH 2010 papers*, ACM, New York, NY, USA, SIGGRAPH '10, 124:1–124:11.

MARSELLA, S., XU, Y., FENG, A., LHOMMET, M., SCHERER, S., AND SHAPIRO, A. 2013. Automated character performance from audio. In *Symposium on Computer Animation, under review*, SCA '13.

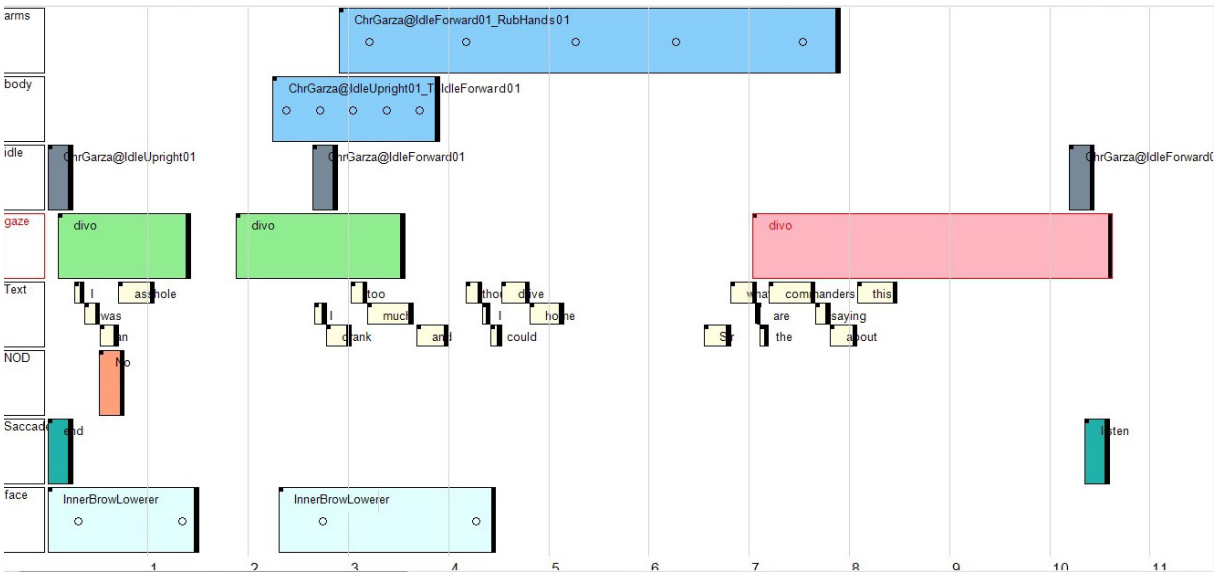

Figure 6: *Our animation method automatically generates a high-level description of the desired character behavior using an XML-based decription called BML (Behavior Markup Language). The BML dictates blocks of behaviors that include timing and other information for various behaviors, such as gestures, head movements, gazes and so forth. These behaviors can be edited by: changing the timing of the behaviors, changing behavior options, adding or removing behaviors.*

NEFF, M., KIPP, M., ALBRECHT, I., AND SEIDEL, H.-P. 2008. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics 27*, 1, 5.

NIEWIADOMSKI, R., BEVACQUA, E., MANCINI, M., AND PELACHAUD, C. 2009. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '09, 1399–1400.

SCHERER, S., KANE, J., GOBL, C., AND SCHWENKER, F. 2013. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language 27*, 1, 263–287.

SFM, 2013. Source filmmaker, http://www.sourcefilmmaker.com/.

SHAPIRO, A. 2011. Building a character animation system. In *Motion in Games*, J. Allbeck and P. Faloutsos, Eds., vol. 7060 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 98–109.

XTRANORMAL, 2013. Xtranormal, http://www.xtranormal.com/.

XU, Y., FENG, A. W., AND SHAPIRO, A. 2013. A simple method for high quality artist-driven lip syncing. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, ACM, New York, NY, USA, I3D '13, 181–181.