BEAT: the Behavior Expression Animation Toolkit*

Justine Cassell[1], Hannes Högni Vilhjálmsson[2], and Timothy Bickmore[3]

[1] MIT Media Laboratory, 20 Ames St., E15-315 Cambridge, MA 02139, USA
justine@media.mit.edu
[2] MIT Media Laboratory, 20 Ames St., E15-320R Cambridge, MA 02139, USA
hannes@media.mit.edu
[3] MIT Media Laboratory, 20 Ames St., E15-320Q Cambridge, MA 02139, USA
bickmore@media.mit.edu

Summary. The Behavior Expression Animation Toolkit (BEAT) allows animators to input typed text that they wish to be spoken by an animated human figure, and to obtain as output appropriate and synchronized non-verbal behaviors and synthesized speech in a form that can be sent to a number of different animation systems. The non-verbal behaviors are assigned on the basis of actual linguistic and contextual analysis of the typed text, relying on rules derived from extensive research into human conversational behavior. The toolkit is extensible, so that new rules can be quickly added. It is designed to plug into larger systems that may also assign personality profiles, motion characteristics, scene constraints, or the animation styles of particular animators.

1 Introduction

The association between speech and other communicative behaviors poses particular challenges to procedural character animation techniques. Increasing numbers of procedural animation systems are capable of generating extremely realistic movement, hand gestures, and facial expressions in silent characters. However, when voice is called for, issues of synchronization and appropriateness render disfluent otherwise more than adequate techniques. And yet there are many cases where we may want to animate a speaking character. While

* This chapter is a reprint from the Proceedings of SIGGRAPH'01, August 12–17, Los Angeles, CA (ACM Press 2001), pp. 477–486. The chapter has been adapted in style for consistency.

H. Prendinger et al. (eds.), *Life-Like Characters*

spontaneous gesturing and facial movement occurs naturally and effortlessly in our daily conversational activity, when forced to think about such associations between non-verbal behaviors and words in explicit terms a trained eye is called for. For example, untrained animators, and some autonomous animated interfaces, often generate a pointing gesture toward the listener when a speaking character says "you". ("If you want to come with me, get your coat on.") A point of this sort, however, never occurs in life (try it yourself and you will see that only if "you" is being contrasted with somebody else might a pointing gesture occur) and, what is much worse, makes an animated speaking character seem stilted, as if speaking a language not its own. In fact, for this reason, many animators rely on video footage of actors reciting the text, for reference or rotoscoping, or more recently, rely on motion captured data to drive speaking characters. These are expensive methods that may involve a whole crew of people in addition to the expert animator. This may be worth doing for characters that play a central role on the screen, but is not as justified for a crowd of extras.

In some cases, we may not even have the opportunity to handcraft or capture the animation. Embodied conversational agents as interfaces to web content, animated non-player characters in interactive role playing games, and animated avatars in online chat environments all demand some kind of procedural animation. Although we may have access to a database of all the phrases a character can utter, we do not necessarily know in what context the words may end up being said and may therefore not be able to link the speech to appropriate context-sensitive non-verbal behaviors beforehand.

BEAT allows one to animate a human-like body using just text as input. It uses linguistic and contextual information contained in the text to control the movements of the hands, arms, and face, and the intonation of the voice. The mapping from text to facial, intonational, and body gestures is contained in a set of rules derived from the state of the art in non-verbal conversational behavior research. Importantly, the system is extremely tunable, allowing animators to insert rules of their own concerning personality, movement characteristics, and other features that are realized in the final animation. Thus, in the same way as Text-to-Speech (TTS) systems realize written text in spoken language, BEAT realizes written text in embodied expressive behaviors. And, in the same way as TTS systems allow experienced users to tweak intonation, pause-length, and other speech parameters, BEAT allows animators to write particular gestures, define new behaviors, and tweak the features of movement.

The next section gives some background to the motivation for BEAT. Section 3 describes related work. Section 4 walks the reader through the implemented system, including the methodology of text annotation, selection of non-verbal behaviors, and synchronization. An extended example is covered in Sect. 5. Section 6 presents our conclusions and describes some directions for future work.