Coordinating Vocal and Visual Parameters for 3D Virtual Agents

Catherine Pelachaud
Dept. of Computer Science and Systems
University of Rome "La Sapienza"
pelachau@graphics.cis.upenn.edu

Scott Prevost
Dept. of Computer and Information Science
University of Pennsylvania
prevost@linc.cis.upenn.edu

Abstract

This paper presents an implemented system for automatically producing prosodically appropriate speech and corresponding facial expressions for animated, three-dimensional agents that respond to simple database queries in a 3D virtual environment. Unlike previous text-to-facial animation approaches, the system described here produces synthesized speech and facial animations entirely from scratch, starting with semantic representations of the message to be conveyed, which are based in turn on a discourse model and a small database of facts about the modeled world.

1 Introduction

As research on the simulation of autonomous virtual human agents progresses, two major issues in human-machine interaction must be addressed. First, proper intonation is necessary for conveying the information structure of utterances with respect to the underlying discourse structure, expressing important distinctions of contrast and focus ([27, 24, 25]). Realistic facial expressions and lip movements help in providing relevant information about discourse structure, turn-taking protocols and speaker attitudes ([8, 9, 18]). Moreover, in a face-to-face conversation, facial displays play an important communicative role.

Simulating this communicative role for animation requires symbolic specification of the semantics and pragmatics of movements. Faces change expressions continuously, and many of these changes are synchronized with what is going on in concurrent conversation. Facial expressions are linked to the content of speech (scrunching one's nose when talking about something unpleasant) as well as affect (smiling when remembering a happy event). They can replace sequences of words (e.g. "the food was [wrinkle nose, stick out tongue]") as well as accompany them [9], and they can serve to help disambiguate what is being said when the acoustic signal is degraded. They do not occur randomly but rather are synchronized to one's own speech, or to the speech of others [6, 15]. It is therefore important that the specification of facial expressions takes many different levels of organization into account. We propose that integrating models for generating proper intonation and facial expressions will improve the intelligibility and naturalness of utterances produced by meaning-to-speech systems as well as by more elaborate systems involving virtual animated human agents (e.g. [3]).

The intonation generation model is based on Combinatory Categorial Grammar (CCG – cf. [27]), a formalism which easily integrates the notions of syntactic constituency, prosodic phrasing and information structure. Based on the CCG grammar, a simple discourse model and a small knowledge base represented in Prolog, the system produces spoken responses to database queries with appropriate intonation. Given the precise timings for phonemes and intonational phenomena in the speech wave, we produce precise specifications for generating the lip movements and facial expressions for a graphical model of a human head. Results from our current implementation demonstrate the system's ability to generate a variety of intonational possibilities and facial animations for a given sentence depending on the discourse context.

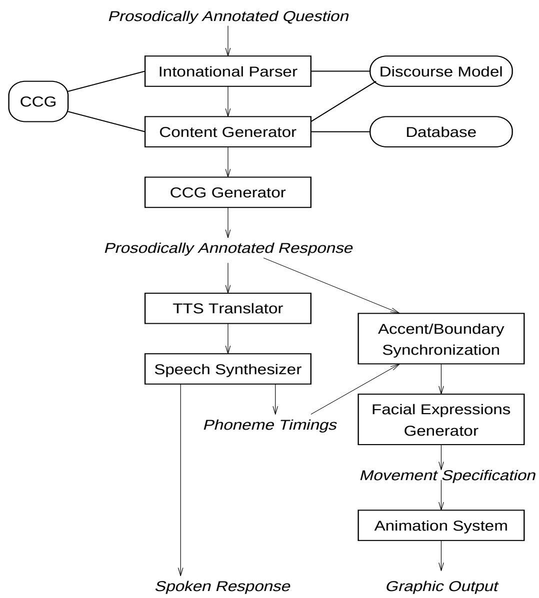

Figure 1: Architecture

Previous work in the area of intonation generation includes studies by Terken ([29]), Houghton and Pearson ([13]), Isard and Pearson ([14]), Davis and Hirschberg (cf. [7, 12]), and Zacharski *et al.* ([31]). Benoit *et al.* ([1]), Brooke ([2]), Cohen *et al.* ([4]), Hill *et al.* ([11]), Lewis *et al.* ([16]) and Terzopoulos *et al.* ([30]) have worked on synchronizing lip movements with speech, producing quite striking results. Takeuchi *et al.* ([28]) implemented a user-interface in which a 3D facial model responds to queries posed by a user. In this system, the generation of the facial expressions accompanying the answer depends on an analysis of the conversational situation and the selection of facial expressions from a database of facial displays.

The system described here expands the work of the aforementioned researchers by linking contextually appropriate intonation with the corresponding facial expressions, and generating the 3D facial animations automatically from semantic, information structural and discourse structural representations [21].

2 The Implementation

Using the CCG theory of prosody outlined in [27, 24, 25], the implemented system undertakes the task of specifying contextually appropriate intonation and facial animation for spoken responses to database queries. The process, which is illustrated in figure 1, begins with a fully segmented and prosodically annotated representation of a spoken query, as shown in example (1), which involves a simple database of facts about stereo components. The notational system representing the intonation contour in example (1) is an adaptation of the widely used system developed by Pierrehumbert ([23]).[1] For simplicity, we show accented words in capital letters without regard for the different possible types of accents. A simple CCG parser determines the semantics of the question, dividing it into its *theme*, which identifies what the sentence is about, and its *rheme*, which identifies what is important or salient about the theme. We refer to this division of the utterance into theme and rheme as its *information structure*. Certain elements of the theme and rheme may be particularly salient because they are new to the discourse or serve to distinguish among entities or propositions that are already firmly established in the discourse. We say such items are in *focus*, and mark them with the * operator, as shown in examples (2).[2]

(1) I know which components produce MUDDY bass,
but WHICH components produce CLEAN bass?
L+H* LH% H* LL$

(2) Proposition:

$s : \lambda x.component(x) \& produce(x, *clean(bass))$

Theme:

$s : \lambda x.component(x) \& produce(x, *clean(bass))/$
$(s : produce(x, *clean(bass)) \backslash np : x)$

Rheme:

$s : produce(x, *clean(bass)) \backslash np : x$

The content generation module has the task of determining the semantics and information structure of the response, marking focused items based on the contrastive stress algorithm described in [25]. For the question given in (1), the strategic generator produces the representation for the response shown in example (3), where the appropriate theme can be paraphrased as "what produces clean bass", the appropriate rheme as "amplifiers", and where the context includes alternative components and audio qualities.

(3) Proposition:

$s : produce(*amplifiers, *clean(bass))$

Theme:

$s : produce(x, *clean(bass)) \backslash np : x$

Rheme:

$np : *amplifiers$

[1]The L+H* and H* markings represent different types of pitch accents in the fundamental frequency contour. The LH% and LL$ markings represent prosodic boundaries. For a brief explanation of the Pierrehumbert-style markings, see [26].

[2]A full explanation of the semantic and syntactic representation in (2) is beyond the scope of this paper. The interested reader should refer to [27, 26].

Using the output of the content generator, the CCG generation module (described in [24]) produces a string of words and Pierrehumbert-style markings representing the response, as shown in example (4).

(4) AMPLIFIERS produce CLEAN bass.
H* L L+H* LH$

The final aspect of speech generation involves translating such a string into a form usable by a suitable speech synthesizer. The current implementation uses the Bell Laboratories TTS system [17] as a post-processor to synthesize the speech wave and produce precise timing specifications for phonemes. The duration specifications are then automatically annotated with pitch accent peaks and intonational boundaries in preparation for processing by the facial expression rules (see also [3]).

Most facial animation systems use the Facial Action Coding System (*FACS*), developed by Ekman and Friesen [10], to annotate facial action. The system describes the visible muscular action based on anatomical studies, using basic elements called action units (*AU*), which refer to the contraction of one muscle or a group of related muscles. A facial expression is described as a set of *AU*s.

Certain facial expressions, which serve *informational structural* functions, accompany the flow of speech and are synchronized at the verbal level. Facial movements (such as raising the eyebrows or blinking while saying "AMPLIFIERS produce CLEAN bass") can appear during accented syllables or pauses. These function are based on the following determinants: conversational signals, punctuators and manipulators. *Conversational signals* correspond to movement occurring on accented or emphatic items to clarify or support what is being said. These can be eyebrow movements (the most commonly used facial expression), head nods, or blinks. *Punctuators* are movements which occur on pauses, reducing the ambiguity of the speech by grouping or separating sequences of words into discrete unit phrases [5]. Slow head movement, blinks, or a smile can accompany a pause. *Manipulators* correspond to biologically necessary functions like blinking to wet the eyes.

As we have seen, a facial expression can have a variety of different meanings (e.g. accentuating an element, punctuating a pause). We propose a high level programming language to describe them, amounting to a formal notation for the different clusterings of facial expressions. Indeed, rather than using a set of *AU*s to specify facial expressions in terms of intonational features in speech, it is more convenient to express them at a higher level, directly denoting their function. These operations are then mapped onto sequences of *AU*s so that we are able to model different facial "styles", in the sense that people differ in their way of emphasizing a word and in the number of facial displays they use. For example, Ekman [9] found that most people use raised eyebrows to accompany an accent while the actor Woody Allen uses eyebrow positions (inner and downward) which generally imply sadness.

Our algorithms incorporate synchrony ([6]), create coarticulation effects, emotional signals, and eye and head movements ([19, 20]). The facial animation system scans the input utterances and computes the associated movements for the lips, the conversational signals and the punctuators. Conversational signals start and end with the accented word. For instance, on *amplifier*, the brow starts raising on 'a', remains raised until the end of the word, and ends raising on ‘r’. On the other hand, the punctuator signals, such as smiling, coincide with pauses. Blinking is synchronized at the phoneme level, due to biological necessity, accentuation or pausing. On *amplifier*, for example, the eyes start closing on ‘a’, remain closed on ‘m’ and start opening on ‘p’.

The computation of the lip shape is done in three passes. First, phonemes, which are characterized by their degree of deformability, are processed one segment at a time using the look-ahead model to search for the proximal deformable segments whose associated lip shapes influence the current segment. For example, in *amplifier* the ‘l’ receives the same lip shape as the following vowel ‘i’—that is, the movement of the ‘i’ begins before the onset of its sound. Second, the spatial properties of muscle contractions are taken into account by adjusting the sequence of contracting muscles when antagonistic movements succeed one another (i.e. movements involving very different lip positions, such as pucker movements versus the extension of the lips). And finally, the temporal properties of muscle contractions are considered by determining whether a muscle has enough time to contract before (or relax after) the surrounding lip shape.

The tongue, although not highly visible, is an important element of distinction between phonemic elements, especially when these elements are not differentiated by their lip shapes. The tongue is composed of 2 parallel surfaces, each of them made of 10 triangles. A tongue shape is defined by varying the tongue parameters, including the length of the edges of the triangles and the angles between each of the edges. Modifying the length of the edges allows for the narrowing, flattening, stretching and/or compression of the tongue, while changing the value of the angles between edges allows the tongue to bend, curve and/or twist. This model is a simplification of [22].

3 Examples

In the examples shown below, the speaker manifests different behaviors depending on whether s/he is asking a question, making a statement, accenting a word or pausing. When asking a question, the speaker raises the eyebrows and looks up slightly to mark the end of the question. When replying, or when turning over the floor to the other person, the speaker turns the head toward the listener. To emphasize a particular word, s/he raises the eyebrows and/or blinks. During the brief pauses at the end of statements and within statements, the speaker blinks and smiles.

(5) I know which amplifier produces clean BASS,

but which amplifier produces clean TREBLE?
L+H* LH% H* LL$

The BRITISH amplifier produces clean TREBLE.
H* L L+H* LH$

(6) I know which British component produces MUDDY treble,

but which British component produces CLEAN treble?
L+H* LH% H* LL$

The British AMPLIFIER produces CLEAN treble.
H* L L+H* LH$

In utterance (5), the word *British* is accented and accompanied by a raised eyebrow, which indicates a conversational signal denoting contrast. In utterance (6), on the other hand, the word *amplifier* is accented and marked by the action of the eyebrows and a blink (see figure 2 in Appendix). The same argument differentiates the appearance of the movement on the word *treble* in (5) and the word *clean* in (6). Moreover, a punctuating blink marks the end of (6), starting on the pause after the word *treble*. In (5) a blink coincides with the accented word *treble* (as a conversational signal) and with the pause marking the end of the utterance (as a punctuator), resulting in two blinks emitted in succession at the end of the utterance. In both examples, the pause between the two intonational phrases '*the British amplifier*' and '*produces clean treble*', is accompanied by movement of the eyebrows.

Figure 2: 'amplifier'

4 Conclusions

The system described above produces quite sharp and natural-sounding distinctions of intonation contour, as well as visually distinct facial animations, for minimal pairs of queries and responses generated automatically from a discourse model and a simple knowledge base. The examples in the previous section (and others presented at the workshop) illustrate the system's capabilities and provide a sound basis for exploring the role of intonation and facial expressions in a 3D virtual environment. Future areas of research include evaluating results and exploring the relevance of our current system to large scale animation systems involving autonomous virtual human agents (cf. [3]).

5 Acknowledgments

We would like to thank particularly Dr. Norman I. Badler and Dr. Mark Steedman for their very useful comments. We are grateful to AT&T Bell Laboratories for allowing us access to the TTS speech synthesizer, and to Mark Beutnagel, Julia Hirschberg, and Richard Sproat for patient advice on its use. The usual disclaimers apply. The research was supported in part by NSF grant nos. IRI90-18513, IRI90-16592, IRI91-17110 and CISE IIP-CDA-88-22719, DARPA grant no. N00014-90-J-1863, and ARO grant no. DAAL03-89-C0031.

6 References

[1] C. Benoit: Why synthesize talking faces? In: Proceedings of the ESCA Workshop on Speech Synthesis, pages 253–256, Autrans, 1990.

[2] N.M. Brooke: Computer graphics synthesis of talking faces. In: Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 1990.

[3] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone: Animated conversation: Rule based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In: SIGGRAPH'94, 1994.

[4] M. M. Cohen and D. W. Massaro: Modeling coarticulation in synthetic visual speech. In: D. Thalmann N. Magnenat-Thalmann (eds.): Computer Animation '93. Springer-Verlag, 1993.

[5] G. Collier: Emotional expression. Lawrence Erlbaum Associates, 1985.

[6] W.S. Condon and W.D. Osgton: Speech and body motion synchrony of the speaker-hearer. In: D.H. Horton and J.J. Jenkins (eds.): The perception of Language, pages 150–184. Academic Press, 1971.

[7] J. Davis and J. Hirschberg: Assigning intonational features in synthesized spoken discourse. In: Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics, pages 187–193, Buffalo, 1988.

[8] S. Duncan: Some signals and rules for taking speaking turns in conversations. In Weitz (ed.): Nonverbal Communication. Oxford University Press, 1974.

[9] P. Ekman: About brows: emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (eds.): Human ethology: claims and limits of a new discipline: contributions to the Colloquium, pages 169–248. Cambridge University Press, Cambridge, England; New-York, 1979.

[10] P. Ekman and W. Friesen: Facial action coding system. Consulting Psychologists Press, 1978.

[11] D.R. Hill, A. Pearce, and B. Wyvill: Animating speech: an automated approach using speech synthesised by rules. The Visual Computer, 3:277–289, 1988.
[12] J. Hirschberg: Accent and discourse context: Assigning pitch accent in synthetic speech. In: Proceedings of AAAI: 1990, pages 952–957, 1990.
[13] G. Houghton and M. Pearson: The production of spoken dialogue. In: M. Zock and G. Sabah (eds.): Advances in Natural Language Generation: An Interdisciplinary Perspective, Vol. 1. Pinter Publishers, London, 1988.
[14] S. Isard and M. Pearson: A repertoire of British English intonation contours for synthetic speech. In: Proceeding of Speech '88, 7th FASE Symposium, pages 1223-1240, Edinburgh, 1988.
[15] A. Kendon: Some relationships between body motion and Speech. In: A.W. Siegman and B. Pope (eds.): Studies in Dyadic Communication, pages 177-210, 1972.
[16] J.P. Lewis and F.I. Parke: Automated lip-synch and speech synthesis for character animation. CHI + GI, pages 143–147, 1987.
[17] M. Liberman and A. L. Buchsbaum: Structure and usage of current Bell Labs text to speech programs. Technical Memorandum TM 11225-850731-11, AT&T Bell Laboratories, 1985.
[18] D.W. Massaro: Speech perception by ear and eye: a paradigm for psychological inquiry. Cambridge University Press, 1989.
[19] C. Pelachaud, N.I. Badler, and M. Steedman: Linguistic issues in facial animation. In: N. Magnenat-Thalmann and D. Thalmann (eds.): Computer Animation '91, pages 15–30. Springer-Verlag, 1991.
[20] C. Pelachaud, M.L. Viaud, and H. Yahia: Rule-structured facial animation system. In: IJCAI 93, 1993.
[21] C. Pelachaud and S. Prevost: Sight and sound: generating facial expressions and spoken intonation from context. In: Proceedings of the Second ESCA Workshop on Speech Synthesis, New Paltz, NY, 1994.
[22] C. Pelachaud, C.W.A.M van Overveld and C. Seah: Modeling and animating the human tongue during speech production. In: Computer Animation '94, Geneva, May, 1994.
[23] J. Pierrehumbert: The phonology and phonetics of English intonation. PhD Dissertation, MIT (Dist. by Indiana University Linguistics Club, Bloomington, IN), 1980.
[24] S. Prevost and M. Steedman: Generating contextually appropriate intonation. In: Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics, pages 332–340, Utrecht, 1993.
[25] S. Prevost and M. Steedman: Using context to specify intonation in speech synthesis. In: Proceedings of the 3rd European Conference of Speech Communication and Technology (EUROSPEECH), pages 2103–2106, Berlin, 1993.
[26] S. Prevost and M. Steedman: Specifying intonation from context for speech synthesis. Speech Communication, 15(1-2), pages 139–153, 1994.
[27] M. Steedman: Structure and intonation. Language, pages 260–296, 1991.
[28] A. Takeuchi and K. Nagao: Communicative facial displays as a new conversational modality. In: ACM/IFIP INTERCHI '93, Amsterdam, 1993.
[29] J. Terken: The distribution of accents in instructions as a function of discourse structure. Language and Structure, 27:269–289, 1984.

[30] D. Terzopoulos and K. Waters: Techniques for realistic facial modelling and animation. In: N. Magnenat-Thalmann and D. Thalmann (eds.): Computer Animation '91, pages 45–58. Springer-Verlag, 1991.

[31] R. Zacharski, A.I.C. Monaghan, D.R. Ladd, and J. Delin: BRIDGE: Basic research on intonation in dialogue generation. Technical report, HCRC: University of Edinburgh, 1993. Unpublished manuscript.