September 1994

Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context

Scott Prevost
University of Pennsylvania

Catherine Pelachaud
University of Pennsylvania

Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context

Abstract
This paper presents a model for automatically producing prosodically appropriate speech and corresponding facial expression for agents that respond to simple database queries in a 3D graphical representation of the world. This work addresses two major issues in human-machine interaction. First, proper intonation is necessary for conveying information structure, including important distinctions of contrast and focus. Second, facial expressions and lip movements often provide additional information about discourse structure, turn-taking protocols and speaker attitudes.

Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context

Catherine Pelachaud and Scott Prevost
Computer and Information Science, University of Pennsylvania
200 South 33rd Street, Philadelphia, PA 19104-6389, USA
Internet contact: prevost@linc.cis.upenn.edu

Preliminary draft accepted for presentation at the Second ESCA/IEEE Workshop on Speech Synthesis

1 Introduction

This paper presents a model for automatically producing prosodically appropriate speech and corresponding facial expression for agents that respond to simple database queries in a 3D graphical representation of the world. This work addresses two major issues in human-machine interaction. First, proper intonation is necessary for conveying information structure, including important distinctions of contrast and focus. Second, facial expressions and lip movements often provide additional information about discourse structure, turn-taking protocols and speaker attitudes ([7], [8], [14], [15]).

The intonation generation model is based on Combinatory Categorial Grammar (CCG – cf. [20]), 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 domain-independent knowledge base, the system produces spoken responses to database queries with appropriate intonation. Given the 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.

Previous work in the area of intonation generation includes studies by Terken

([21]), Houghton, Isard and Pearson (cf. [11]), Davis and Hirschberg (cf. [6], [10]), and Zacharski et al. ([23]). Benoit et al. ([1]), Brooke ([2]), Cohen et al. ([4]), Hill et al. ([9]), Lewis et al. ([12]) and Terzopoulos et al. ([22]) have worked on lip synchronization with speech.

2 The Implementation

The present paper presents an implemented system that applies the CCG theory of prosody outlined in [20], [18] and [19] to the the task of specifying contextually appropriate intonation and facial animation for spoken responses to database queries. The process begins with a fully segmented and prosodically annotated representation of a spoken query, as shown in example (1). We employ a simple bottom-up shift-reduce parser to identify the semantics of the question, dividing it into a topic or "theme" and a comment or "rheme", and marking "focused" items with *, as shown in (2).

(1) I know what the CAT scan is for,
but WHICH condition does URINALYSIS address?
L+H* LH% H* LL$

(2) Proposition: `s :lambda(x, condition(x) & address(*urinalysis, x))`
Theme: `s:lambda(x, condition(x) & address(*urinalysis, x))/`
`(s:address(*urinalysis, x)/np:x)`
Rheme: `s:address(*urinalysis, x)/np:x`

The strategic generation module has the task of determining the semantics and information structure of the response, marking focused items based on an algorithm described in [19]. For the question given in (1), the strategic generator produces the following representation for the response, because the theme is "what urinalysis addresses", the rheme is "hematuria", and the context includes alternative conditions and treatments:

(3) Proposition: `s:address(*urinalysis, *hematuria)`
Theme: `s:address(*urinalysis, x)/np:x`
Rheme: `np:*hematuria`

From the output of the strategic generator, the tactical generation module (described in [18]) produces a string of words and Pierrehumbert-style markings representing the response, as shown in (4).

(4) URINALYSIS addresses HEMATURIA
L+H* LH% H* LL$

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 [13] as a post-processor to synthesize the speech wave and produce precise timing specifications for phonemes. The duration specifications are then annotated with pitch accent peaks and intonational boundaries before being sent to the animation system for processing ([3]).

Starting from a functional group (lip shapes, conversational signal, punctuator, regulator or manipulator), we offer algorithms which incorporate synchrony [5], create coarticulation effects, emotional signals, and eye and head movements ([16], [17]). Our rules generate automatically the facial actions corresponding to an input utterance. A conversational signal (movements occurring on accents, like raising of eyebrow) starts and ends with the accented word; while punctuator signals (such as smiling) coincide with pauses. Blinking is synchronized at the phoneme level. Head nods and shakes appear on accent and pause. The head of the speaker turns away from the listener at the beginning of a speaking turn and turns toward to the listener at end of a speaking turn to signal a change of turn.

The computation of the lip shape is done in three passes. Phonemes are characterized by their degree of deformability. For each deformable segment, the program looks for the nearby segment whose associated lip shapes influence it using the look-ahead model; The properties of muscle contractions are taken into account in two ways: spatially, by adjusting the sequence of contracting muscles if antagonist movements (i.e. movements which show very different lip positions like pucker movements versus extension of the lips) succeed each other; and temporally by noticing if a muscle has enough time to contract (respectively relax) before (respectively after) the surrounding lip shape. Those two muscle constraints act on the final computation of the lip shapes.

3 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. The examples in the full paper illustrate the system's capability for producing appropriately different audio and visual output for a single string of words under the control of differing discourse contexts. We believe the system provides a sound basis for exploring the role of prosody and facial expressions in human-machine interactions, particularly those involving autonomous virtual human agents.

References

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

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

[3] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, Chin Seah, and Matthew Stone. Animated conversation:rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *ACM SIGGRAPH 94*, 1994. submitted.

[4] Michael M. Cohen and Dominic W. Massaro. Modeling coarticulation in synthetic visual speech. In D. Thalmann N. Magnenat-Thalmann, editor, *Computer Animation '93*. Springer-Verlag, 1993.

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

[6] James Davis and Julia 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.

[7] S. Duncan. Some signals and rules for taking speaking turns in conversations. In Weitz, editor, *Nonverbal Communication*. Oxford University Press, 1974.

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

[9] 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.

[10] Julia Hirschberg. Accent and discourse context: Assigning pitch accent in synthetic speech. In *Proceedings of AAAI: 1990*, page ??, 1990.

[11] George Houghton. *The Production of Language in Dialogue: a Computational Model*. PhD thesis, University of Sussex, 1986.

[12] J.P. Lewis and F.I. Parke. Automated lip-synch and speech synthesis for character animation. *CHI + GI*, pages 143–147, 1987.

[13] Mark 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.

[14] D.W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Cambridge University Press, 1989.

[15] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[16] C. Pelachaud, N.I. Badler, and M. Steedman. Linguistic issues in facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 15–30. Springer-Verlag, 1991.

[17] C. Pelachaud, M.L. Viaud, and H. Yahia. Rule-structured facial animation system. In *IJCAI 93*, 1993.

[18] Scott Prevost and Mark 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.

[19] Scott Prevost and Mark Steedman. Using context to specify intonation in speech synthesis. In *Proceedings of EuroSpeech 93*, Berlin, 1993. To appear.

[20] Mark Steedman. Structure and intonation. *Language*, pages 260–296, 1991.

[21] Jacques Terken. The distribution of accents in instructions as a function of discourse structure. *Language and Structure*, 27:269–289, 1984.

[22] D. Terzopoulos and K. Waters. Techniques for realistic facial modelling and animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 45–58. Springer-Verlag, 1991.

[23] 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.