September 1994

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

Scott Prevost
University of Pennsylvania

Catherine Pelachaud
University of Pennsylvania

Follow this and additional works at: http://repository.upenn.edu/hms

Recommended Citation


This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the University of Pennsylvania's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/hms/40
For more information, please contact libraryrepository@pobox.upenn.edu.
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.

Comments

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the University of Pennsylvania's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this document, you agree to all provisions of the copyright laws protecting it.
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 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


