Artefact Specification

CreST Voice Expressivity & Emotion Group

NOTE TO READERS: This document is INCOMPLETE, as I want to get it out to the VEEG group in as early and immature a state as possible to allow maximum collaboration. I believe that this reflects most of what we all discussed in York, as well as the comments you have sent me since. Do not be alarmed at seeming complexity, confusion, lack of detail, or contradictions—we're still in the exploration phase. The goal is a detailed design ready to start coding before the spring meeting.

1. How to Use These Documents

All documents are distributed in PDF format to avoid version and operating system incompatibilities. If you want the originals for direct editing, just ask. By default, all documents are formatted for A4 paper. If you want an 8.5" x 11" version, just ask. *n.n* represents the version number, be sure to have the correct (latest) version for each document. Different documents may be at different version levels.

1.1 Documents Summary

Title (.PDF)	Description	Reader	Editable Format
CreST-VEEG-ArtifactSpec-v*n.n*	This document—explicates all of the other documents.	All readers will need to refer to this document repeatedly. This is the main reference document.	Microsoft Word 2003 for Windows XP
CreST-StateMachine-v*n.n*	State-transition diagrams and supporting flowcharts	Programmers will use this for software details. Other readers may find some illustrations helpful for high-level path-tracing.	Microsoft Visio 2003 for Windows XP
CreST-StateTables-v*n.n*	Text to be spoken and recognized, as well as grammars and parameters that control chatbot behavior.	Poets, script writers, and other wordsmiths will use these tables to read and write dialogues.	Microsoft Word 2003 for Windows XP (landscape)

1.2 State-Transition Diagrams

Step-by-step discussions will appear in chapter 3 of this document. Chapter 3 is currently empty, awaiting group contribution and design. The state-transition diagrams should be self-explanatory for technical people, and readable for non-technical people. We'll schedule a meeting in January to go over them in some detail, in the meantime, e-mail me with comments or questions.

2. Physical Appearance

2.1 Heads/Faces/Puppets 

There are four entities, each represented by some sort of physical embodiment. At the October workshop, we considered Styrofoam wig stands, hand-made doll heads, or even graphical faces on a computer screen. The four entities are symbolized in Figure 1 by the blue circles. These can (and should) be designed by sculpture, set design, industrial design, or other visual arts skills within the CreST network.

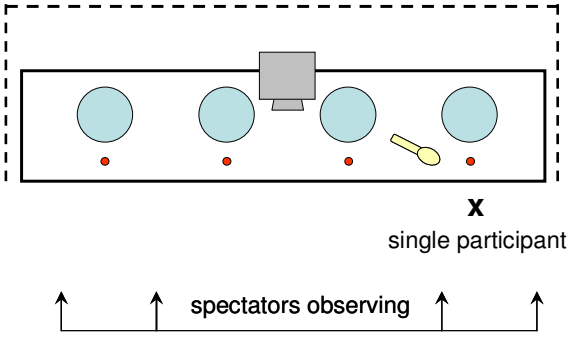

Figure 1—Example Schematic of Artefact

The four entities may be arranged side-by-side on a table as shown in Figure 1. The dotted line shows table-skirting or drapery to prevent spectators from going beside or behind the exhibit. The artifact may be one of several exhibits in an arcade, or may be part of an installation in an art museum or gallery, as discussed at CreST meetings.

2.2 Facial Expression (no icon)

Each entity has a "face" of some kind. The face may be animated in several ways:

- LED lights that represent facial features;
- Animation of some kind on the laptop display screen;
- Overhead illumination (like an interrogation lamp);
- Movement of lips or jaw; and/or,
- Other visual expressions.

2.3 Press-to-Select Buttons

In front of each entity is a large button, labeled "Press to Select." There are four such buttons. Pressing any button "activates" its corresponding entity—that is, triggers the facial expression through illumination or motion. Only one entity at a time may be active. Pressing any button therefore deactivates other entities, and activates the selected entity.

2.4 Audio System

An audio input and output system supports spoken conversation. The output (mono) system consists of a loudspeaker—placed at any convenient location—that is connected to the speaker output jack of the laptop. Text-to-speech voice signals are presented via this loudspeaker to the ambient space, causing the active entity to appear to be talking. The input system consists of a wireless microphone—which the single participant may pick up and carry freely around—that is connected to the mic input on the laptop. Speech recognition software on the laptop receives speech input and recognizes it, giving the appearance that the human participant is talking to and being understood by the active entity. Alternation between talking and recognizing constitutes a an "artificial conversation".

In the event of difficulties with acoustic/phonetic models, acoustical feedback, or other audio problems, there are alternatives to the audio system:

- telephone handset;
- closed headphones; or,
- other alternatives TBD.

2.5 Single Participant x

One person at a time may operate the exhibit. This may be any one of the spectators, monitored or not. The single participant may also be a CreST actor trained in the exhibit dialogues.

2.6 Spectators

Spectators see the facial expressions, hear the human speech produced by the single participant, and hear the artificial (TTS) speech produced in response. They are thus observing the artificial conversation.

2.7 Alternative Physical Arrangements

An alternative physical arrangement is shown in Figure 2. The entities are arranged in a circle, facing inward. In such an arrangement, the single participant may sit in the middle, perhaps on a rotating bar stool, for easy access to all four entities. Spectators may surround the exhibit.

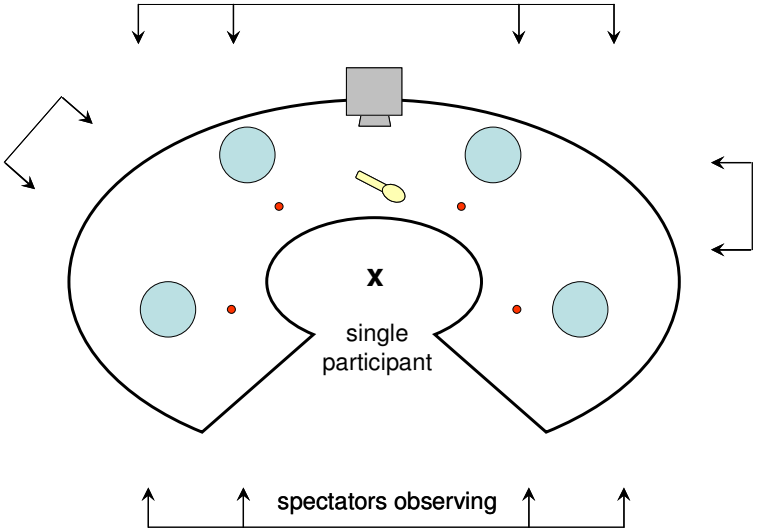

Figure 2—Alternative Circular Layout

A third alternative is shown in Figure 3. Each entity is on a separate freestanding pedestal, with interconnecting wires hidden away. The audience roams freely among the pedestals. In this context, there may be multiple patrons competing to be the single participant, requiring minor changes in the behavioral design of the artefact.

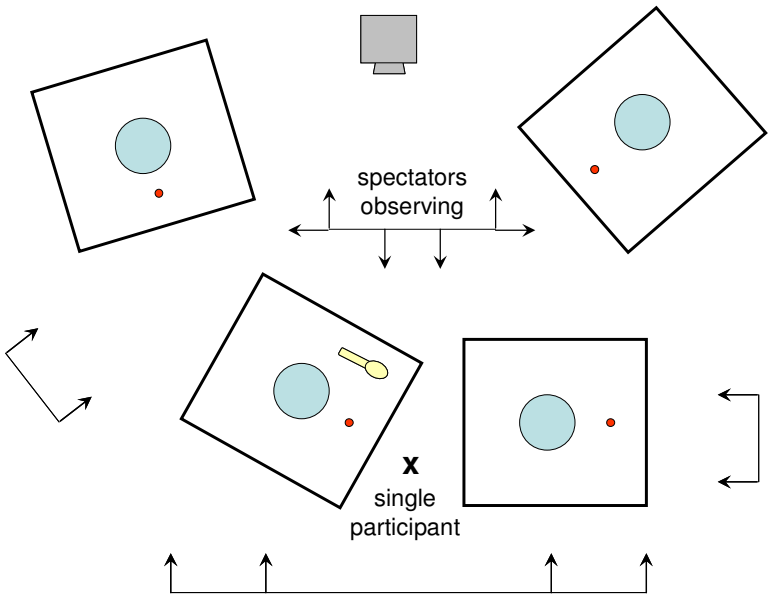

Figure 3—Pedestal Arrangement

3. State-Transition Diagrams

This chapter will explicate all diagrams. This will be helpful later when details dominate the project, but is less important now. 3.1 shows a complete explication of the first (High-Level Main) diagram, which should be enough to provide an overview.

3.1 High-Level Main

Refer to the current version of the state-transition diagrams, page High-Level Main.

1. The START bubble shows the start of the program.
2. Initialization is described later in this chapter. There are two exits, “done” and “abend”. The “else” path indicates that all other input is ignored.
3. Successful initialization causes a transition to the quiescent state. In this state, there is no user (single participant). The exhibit is engaged in a repetitive behavior aimed at attracting attention to itself. This behavior entails changing the facial expressions and voices while speaking.
4. As long as there is no input from a single participant, the exhibit remains in the quiescent state.
5. The single participant starts a conversation by pressing any of the four “press to select” buttons.
6. Pressing a button causes a transition to the sentient interaction state. The single participant has started a conversation.
7. As long as the single participant follows the planned path for a given dialogue—and as long as there are no false speech recognition events—the sentient dialogue proceeds. Sentient dialogues are rational and calm, with little variation in the voice parameters. This “affective narrowing” increases the longer the sentient state remains active.
8. If the user stops interacting, then a timeout from the ASR triggers a transition back to the quiescent state. The dialogue has been abandoned in mid-conversation, and the same or a new participant must now press a button to start a new one.
9. Errors from the ASR trigger a transition to the rubbish state. The single participant does not know which state is active, but the rubbish state generates random dialogues with greater ambiguity. At no point does the exhibit declare that any kind of “error” condition exists. The only difference to the user is that some (rubbish) conversations cause the machine to diverge from “normalcy”, while other (sentient) conversations seem to make more sense.
10. In the rubbish state, conversations are confusing and/or surreal, and voice parameter changes are more extreme. The machine becomes “agitated” or “emotional” in this state. This “affective divergence” increases the longer the rubbish state remains active.
11. If the user stops interacting, then a timeout from the ASR triggers a transition back to the quiescent state. The dialogue has been abandoned in mid-conversation, and the same or a new participant must now press a button to start a new one.
12. If the single participant reenters a planned path, the event triggers a return to the sentient interaction state. The “affective narrowing” that is the signature of a sentient dialogue may take awhile to occur, however, as voice parameters will change slowly and incrementally from their most recent settings in the rubbish state.
13. As long as the single participant detours from any planned dialogue path—as detected by timeouts, no-match conditions, low ASR confidence, and other spurious speech recognition events—the rubbish dialogue proceeds. Voice parameters may reach their extremes.
14. The user has an assigned task—to match, or “align” the voices with the heads. If this task is accomplished, the condition triggers a transition to the Final Goal state.
15. In the Final Goal state, the exhibit has a final conversation with the user.
16. The conversation takes a few turns and must remain sentient to succeed. Any cooperative user should be able to succeed with this conversation.
17. If the conversation remains sentient and the user succeeds at choosing the correct head-voice pair, then a reward of some kind—somewhat akin to the bells and lights of a winning slot machine—indicates

that the game is over. Success triggers a transition to the quiescent state to await a new single participant and a new conversation.

18. Failure to achieve the final goal returns the dialogue to the rubbish state.
19. The program can be stopped from within the quiescent state. Input to stop the program is TBD, but presumably is a key such as Esc, or a key combination entered at the laptop keyboard. The Single Participant never starts or stops the program.
20. If initialization fails, the “abnormal end” path exits the program. The programmer may peruse log files or core dumps at this time. This exit should never happen during a formal performance or exhibit.
21. The program has stopped running.

4. Reading the State Tables

TBD

5. Examples

5.1 Sample Performance Dialogues

Following are some examples of the kind of conversation that might be had between CreST actors and the kiosks.

Kiosk: Talking and listening sort of go hand in hand.

Human: Hand in hand?

Kiosk: First you present these bizarre sounds.

Human: (nods without speaking)

Kiosk: And then you're taking them in.

Human: But it's really about communication, isn't it?

Kiosk: Communication?

Human: I talk, you listen; I listen, you talk.

Kiosk: There's more to listening than recognizing speech.

Human: I'm not sure I agree.

Kiosk: It's odd, isn't it?

Human: What's that?

Kiosk: How fragile a conversation can be.

Human: Yes?

Kiosk: I mean ... one moment, you feel engaged with a mind; it all seems to be making some kind of sense. And then moments later, you're drowning in a sea of doubt.

Human: A sea of doubt?

Kiosk: Do you ever wonder that it's all a hoax?

5.2 Sample Interactive Spectator Dialogues

TBD

6. Appendices

6.1 CreST

The Creative Speech Technology (CreST) network is an interdisciplinary network of contributors to the field of computer speech. It is led by Dr Christopher Newell, University of Hull and Dr Alistair Edwards, University of York. It is funded by an EPSRC[1] grant and is scheduled to run from March 2011 for 2 years. The output is expected to be a public display or performance of some kind, probably in November, 2012.

CreST was proposed and funded under the auspices of the Interdisciplinary and Collaborative Practices (ICP) Cluster[2] at the University of Hull.

The following is quoted from the ICP website:

"The evaluation of synthetic speech is problematic. We need a language of evaluation that is focused neither on intelligibility nor verisimilitude to human speech. The arts may have evolved evaluation criteria that could help address this problem. ... [The idea of] *appropriateness* seems to encapsulate a number of evaluation criteria without constraining imagination and ambition. ... Teasing out the factors in the speech signal that influence the perception of appropriateness [therefore] presents a worthwhile technical problem for the network."[3]

CreST research is inherently interdisciplinary—comprised of artistic, scientific, engineering, and sociological interests. Finding common ground, searching for shared attributes, drawing connections between disciplines, and defining broad and general goals through a collaborative process are deeply ingrained in the network's *raison d'être*.

There are four objectives for the CreST network:

- Support for the emergence of common languages and understandings that facilitate better communication between speech scientists and speech practitioners in the arts. Encourage the emergence of an interdisciplinary ontology for computer speech production;
- To encourage the development of partnerships and networks between speech scientists and speech practitioners in the arts that will lead to creative collaborations and further research proposals;
- To provide a platform for the development of modestly scaled prototypes and artifacts by network members; and,
- To provide heightened public engagement with computer speech research through a touring show demonstrating works developed within the network, special activities for the public throughout the life of the network and an interactive website.

6.2 EPSRC

The Engineering and Physical Sciences Research Council (EPSRC) is the main UK government agency for funding research and training in engineering and the physical sciences, investing more than £850 million a year in a broad range of subjects – from mathematics to materials science, and from information technology to structural engineering. EPSRC is a non-departmental public body funded by the UK government through the Department for Universities, Innovation and Skills.

EPSRC employ around 300 staff in Swindon, and support research into engineering, mathematics, physics, chemistry, materials science, information and communications technologies.

The CreST effort is currently funded by EPSRC. For more information, go to:

http://www.epsrc.ac.uk/Pages/default.aspx

[1] See a brief discussion later in this paper.

[2] http://icpcluster.org

[3] See a more detailed discussion of *appropriateness* later in this paper.

6.3 Appropriateness

For years, we in the speech community have been confronted with subjective questions that relate to voice quality:

- What is the best voice for a given application?
- Whether digitally recorded or synthesized, how should an artificial voice sound to its user? Can it express emotion? Should it?
- Designers of voice applications often speak of personality. What does that mean in reference to an artificial voice? What role does the personality play in the user's interaction with the voice?
- Can a computer be said to have emotion?
- When a voice is "acting"—that is, producing effects aimed at convincing a listener that feelings, emotions, intentions, or personality traits are present when in fact they are not—do listeners respond as though the machine were a real human actor?
- Can acting by an artificial voice be beneficial in an HCI context? An artistic context? A marketing context?
- When an artificial voice is serving as the primary communication medium for a human user—for example people who have lost their voices or do not have adequate motor control for intelligible vocalization—does the voice include identity cues? That is, can the voice be thought of as representing that individual? Is that individual in fact "speaking through" the artificial medium? If so, how does that change the answers to these questions about emotion?

In the course of discussing these issues, CreST members have homed in on the term *appropriateness* as the guiding principle for answering these questions. All of the questions can be rephrased as, "what is the appropriate vocal behavior of an artificial voice?" The answer then becomes, "It depends on the specific set and setting of the user and the context of use." In other words, the answer changes—sometimes dramatically—with changes in context.

CreST has therefore adopted this criterion of *perceived appropriateness* as a working standard against which to measure progress.

End of document
Version 0.07
December 8, 2011
BEB