Roadmaps, Journeys and Destinations
Speculations on the Future of Speech Technology Research

Ronald A. Cole

Center for Spoken Language Research
University of Colorado, Boulder
cole@cslr.colorado.edu

Abstract

This article presents thoughts on the future of speech technology research, and a vision of the near future in which computer interaction is characterized by natural face-to-face conversations with lifelike characters that speak, emote and gesture. A first generation of these perceptive animated interfaces are now under development in a project called the Colorado Literacy Tutor, which uses perceptive animated agents in a computer-based literacy program

1. Introduction

The theme of this Eurospeech03 panel – possible roadmaps for speech technology research – is both timely and important. Speech technology is in a period of great flux. In the past few years, a major speech technology company and at least two major speech research labs in industry have disappeared. These are significant and disturbing events. While the current economy forces some companies to make hard choices, recent decisions to disband major speech labs with brilliant researchers and distinguished histories suggests that corporate decision makers do not believe speech recognition technology will deliver its potential and recoup corporate investment in the near future. And perhaps they are correct: while statistical approaches to speech recognition are elegant and powerful, the cost of creating systems with satisfactory performance for new applications may be too high to produce an acceptable return on investment. Given this state of flux, now is a very good time to consider what might be changed to revitalize speech technology research and produce a new generation of laboratory and commercial systems.

Before considering possible futures for speech technology research, it is important to note that we are not really looking for a roadmap, as a roadmap presents a near-infinite number of routes to a far-off definition. What we need to consider are thoughts about what our destination might be, and what sort of trip we would like to take on the way there. The destination relates to the goals of speech technology research—how speech technology might be used in the future, and what research might be conducted to achieve these goals. And just as important, what is the nature of the journey we choose to undertake to arrive at this destination?

Thinking about a roadmap for the future of speech technology research while writing this article reminded me of my favorite road trip, which occurred in 1988, when I changed jobs drove with my wife Pam from Pittsburgh PA (CMU) to Portland OR (OGI). When planning our trip, the most important decision we made concerned the *nature* of the journey. We decided to optimize fun and new experiences, to take less traveled roads, and to plan interim destinations dynamically while on the road ("Let's go to Glacier Park, we haven't been there.") I recall that once we knew our destination and had decided on the nature of the journey, we didn't look at the actual roadmap to plan our first phase of the trip until we got into our car to start our journey.

Could this be a good strategy for speech technology research: To select a destination, and then work together to decide on the nature of the journey? Below, I will suggest that a useful goal for speech technology research is to achieve Great Communication between people and machines. As to the nature of the journey, I believe that Ron and Pam's Best Road Trip is a good model. That is, once the destination is chosen, the path through the roadmap should be generated dynamically, based on insights gained from past experiences and new ideas. The journey should be characterized by many explorations, and guided by successes and failures during these adventures. And above all, it must be interesting and fun. Stated another way, speech technology research should be guided by ambitious goals, and conducted by independent researchers and research teams who share resources, ideas and knowledge at rest stops along the road (such as Eurospeech).

2. Great Communication: A Guiding Principle

I believe that achieving *Great Communication* between people and machines is a worthy and useful goal for speech technology research. What is great communication? While human communication is infinitely complex and varied, we can identify some common characteristics of great communication experiences.

- Great communication is emotional. Great books, movies, dance performances, learning experiences and conversations engage our emotions. They are often visceral experiences, bringing great joy, sorrow, etc. Great communication experiences speak to the heart as well the mind. *Communication systems of the future must produce emotional experiences.*

- Great communication is immersive—it focuses all of our attention on the communication experience, to

the point where we are fully immersed and unaware of anything else. *Communication systems of the future must produce immersive experiences.*

- Great communication is personal—it affects us deeply during the experience, and can have a lasting impact on our lives. *Communication systems of the future must interact with each user as a unique and special individual.*

Spoken dialogue systems today do not provide great communication experiences. They can be efficient, effective and even satisfying. They can save us time and money. But to achieve their promise, and to help realize human potential, they must evolve from unemotional, impersonal, efficient systems to great communication experiences.

Achieving great communication requires a fundamental change in the way we think about what spoken language systems should do, how they should behave, and how they should be evaluated. Future systems will be evaluated by a new set of measures, based on the outcomes and experiences they produce relative to the objectives of the user. Measures like word recognition accuracy, transaction completion and user satisfaction will be replaced by outcomes such as: Did Bobby learn to read? Is Molly a better writer? Is Todd a better mechanic? Is Dante managing his anger better at home? Is Jessica more self-confident and effective in social situations? Does Monica speak Spanish well enough to work as a teacher in Spain?

3. Perceptive Animated Interfaces

I envision a new generation of human computer interfaces that interact with people like people interact with each other. These interfaces will use intelligent and embodied animated agents to engage users in natural face-to-face conversational interaction to accomplish a wide variety of tasks. An intelligent agent is one that mimics the behaviors of real persons and behaves intelligently in the context of a specific application or task domain. An embodied agent is one that resembles a real person. I call these interfaces of the future *perceptive animated interfaces* [1].

Perceptive animated interfaces are intended to represent a first step toward great communication systems. By using embodied animated agents, we can attempt to develop interfaces that exhibit personality, express emotions, and model signals and cues that characterize nature face-to-face conversational interaction. The invention of interfaces incorporating *virtual humans* [2] is a grand challenge and for speech technology researchers.

Perceptive animated interfaces will be populated with lifelike three-dimensional computer characters that use human language, computer vision and character animation technologies to enable natural face-to-face conversations with users. Animated agents will converse with users just like two people converse with each other— using speech, head nods, eye contact, facial expressions and hand and body gestures. The computer characters will orient to the user, interpret her auditory and visual behaviors to infer her intentions and cognitive state, provide real time feedback while the user is speaking (e.g., to indicate agreement, puzzlement, desire to speak, etc.), engage in interactive turn-taking behaviors, and communicate both linguistic and emotional content using speech, facial expressions and gestures.

The advent of perceptive animated interfaces will enable users to communicate with machines using their natural communication skills, thus producing interactive experiences that are more personal, emotional, meaningful, enjoyable and effective. Natural communication with intelligent animated agents will present new and unprecedented opportunities to individuals, including those who cannot read or type, to learn new skills, communicate more effectively, and increase their participation in the information society.

Perceptive animated agents will change fundamentally the *experience* of communicating with machines by fully engaging user's senses and emotions. An animated agent that behaves as if it perceives the user, understands the user's speech, accurately interprets the user's emotions, and responds in an appropriate and sensitive manner, has the capability to produce intense, immersive and emotional interpersonal experiences.

4. Research Challenges

Building systems that enable face-to-face communication with intelligent animated agents requires a deep understanding of the auditory and visual behaviors that individuals produce and respond to while communicating with each other. Face-to-face conversation is a virtual ballet of auditory and visual behaviors, with the speaker and behavior simultaneously producing and reacting to each other's sounds and movements. While talking, the speaker produces speech annotated by smiles, head nods and other gestures. At the same time, the listener provides simultaneous auditory and visual feedback to the speaker (e.g., "I agree," "I'm puzzled," "I want to speak."). For example, the listener may signal the speaker that she desires to speak; the speaker continues to talk, but acknowledges the nonverbal communication by raising his hand and smiling in a "wait just a moment" gesture. Face-to-face conversation is often characterized by such simultaneous auditory and visual exchanges, in which the sounds of our voices, the visible movements of our articulators, direction of gaze, facial expressions and head and body movements present linguistic information, paralinguistic information (emotions, sarcasm, spatial referents, etc.), and communication about the conversation itself (agreement, turn taking, etc.).

Inventing systems that engage users in natural face-to-face conversational interaction is a challenging task. The system must simultaneously interpret and produce auditory and visual signals in real time while preserving the timing relationships between perception and production appropriate to conversational interaction. The system must interpret the user's auditory and visible speech, eye movements, facial expressions and gestures, since these cues combine to signal the speaker's intent—e.g., a head nod can clarify reference, whereas a shift of gaze can indicate that a response is expected. Paralinguistic information is also critical, since the prosodic contour of the auditory signal or a visual cue such as rolling the eyes may signal that the user is being sarcastic.

The animated agent must also produce accurate, natural, and expressive auditory and visible speech with facial expressions and gestures appropriate to the context of the dialogue, and the goals of the task. Most important, the animated interface must combine perception and production to interact conversationally in real time – while the animated agent is speaking, the system must interpret the user's auditory and visual behaviors to detect agreement, confusion, desire to interrupt, etc., and while the user is speaking, the system must both interpret the user's speech and simultaneously provide auditory and/or visual feedback via the animated character.

To develop lifelike computer characters imbued with unique and credible personalities capable of natural and graceful face-to-face dialogues with users, new research is needed to gain a deeper understanding of the signals and cues exchanged during face-to-face communication, and research is needed to use this knowledge to develop human communication systems that model these behaviors. By applying this knowledge to improved machine perception and generation technologies, research will lead to a new generation of perceptive animated interfaces.

5. First Steps

While perceptive animated interfaces are still science fiction, human communication technologies have matured to the point where it is now possible to conceptualize, develop and test initial system prototypes. I conclude by describing a project, the Colorado Literacy Tutor, which provides a test bed for research and development of perceptive animated interfaces.

5.1. The Colorado Literacy Tutor (CLT)

The Colorado Literacy Tutor is a technology-based literacy program, based on cognitive theory and scientifically based reading research, which aims to improve literacy and student achievement in public schools. The goal of the Colorado Literacy Tutor is to provide computer-based learning tools that will improve student achievement in any subject area by helping students learn to read fluently, to acquire new knowledge through deep understanding of what they read, to make connections to other knowledge and experiences, and to express their ideas concisely and creatively through writing. A second goal is to scale up the program to both state and national levels in the U.S. by providing accessible, inexpensive and effective computer-based learning tools that are easy to use and require little or no learning curve by teachers or students.

A key feature of the Colorado Literacy Tutor is the use of leading edge human communication technologies in learning tasks, as described below. The program is thus an ideal test bed for research and development of perceptive animated agents that integrate auditory and visual behaviors during face-to-face conversational interaction with human learners. The program enables us to evaluate component technologies with real users—students in classrooms—and to evaluate how the integration of these technologies into learning tools incorporating perceptive animated agents affects learning using standardized assessment tools.

The Colorado Literacy Tutor consists of four tightly integrated components: (a) Managed Learning Environment, (b) Foundational Reading Skills Tutors, (c) Interactive Books, and (d) Summary Street comprehension training. In addition, the project devotes significant effort to evaluating learning outcomes and creating a scalable and sustainable program. Here we limit discussion to Interactive Books, the main platform for research and development of perceptive animated agents.

Interactive Book authoring tools enable a wide range of user and system behaviors within Interactive Books, including having the story narrated by one or more animated characters (while controlling their facial expressions and gestures), having conversations with animated characters in structured or mixed-initiative dialogues, having the student read out loud while words are highlighted, clicking on words to have them spoken by the agent, interacting with the agent to sound out the word, having the student respond to questions posed by the agent either by clicking on objects in images or saying or typing responses, and having the student produce typed or spoken summaries which are analyzed for content using language processing techniques.

Read-aloud feedback involves following along as text is read, highlighting the read text, monitoring reading fluency and verifying pronunciation accuracy. Various feedback options are possible, such as display of just the current paragraph being read, highlighting of the read text, and a pointer that follows the current reading position. Read-aloud feedback is obtained by building a language model for the book, getting partial phrases from the speech recognizer as the user is reading, determining the current reading location using the partial phrase and an efficient Viterbi search through the book, and aligning the partial phrase with the book text using a Dynamic Programming search.

5.2. CLT Technology Components

CSLR's Conversational Agent Toolkit (CAT) provides a set of technology modules and tools for researching and developing advanced dialogue systems-- systems that enable completely natural and unconstrained mixed-initiative conversational interaction with users in specific task domains. The component technologies within the Conversational Agent Toolkit are: Face to face conversations between animated characters and students are enabled by communication among a set of technology servers incorporated within CSLR's Conversational Agent Toolkit. These include:

1) Audio Server: The audio server receives signals from the microphone or telephone and sends them to the speech recognizer.

2) Speech Recognizer: Speech recognition plays an integral role in any perceptive animated agent interface. *Sonic* is CSLR's large vocabulary continuous speech recognition system [3], [4]. In addition to large vocabulary speech recognition, the recognizer has been developed to support both keyword / phrase spotting and constrained grammar-based speech recognition. Sonic has been trained on children's speech for use in Interactive Books, described below.

3) Natural Language Parser: We use the Phoenix parser [5] to map the speech recognizer outputs onto a sequence of semantic frames. Because spontaneous speech is often ill formed, and because the recognizer will make recognition errors, Phoenix was designed to be robust to errors in recognition, grammar and fluency.

4) Dialogue Manager: The Dialogue Manager (DM) controls the system's interaction with the user and the application server. It is responsible for deciding what action the system will take at each step in the interaction.

6) Natural Language Generator: The language generation module uses templates to generate words to speak back to the user based on dialogue speech acts. The natural language generator sends the resulting text to the speech generation server for playback to the user.

7) Speech Generation Server: In the Colorado Literacy Tutor, all system prompts are recorded by voice talent, and then synchronized to the visible speech movements of the animated character.

8) *Character Animator:* The character animation module receives a string of symbols (phonemes, animation control commands) with start and end times from the speech generation server, and produces visible speech, facial expressions and hand and body gestures in synchrony with the speech waveform.

Our facial animation system, *CU Animate* [6], is a toolkit designed for research, development, control and real time rendering of 3D animated characters. Ten engaging full-bodied characters are included with the toolkit. Each character has a fully articulated skeletal structure, with sufficient polygon resolution to produce natural animation in regions where precise movements are required, such as lips, tongue and finger joints. Characters produce lifelike visible speech, facial expressions and gestures. CU Animate provides a graphical user interface for designing arbitrary animation sequences. These sequences can be tagged (as icons representing the expression or movement) and inserted into text strings, so that characters will produce the desired speech and gestures while narrating text or conversing with the user.

9) Face Tracker: A face tracking system, developed by Javier Movellan and his colleagues at the Machine Perception Lab at UC San Diego to track faces in real time (at 30 frames per second) under arbitrary illumination conditions and backgrounds (which may include moving objects). The face detector communicates the location of the user's face to the animation server, which, by triangulating between the user, camera and animated agent, allows the animated agent's eyes to track the user.

10) Emotion Monitor: The Emotion Monitor, also developed at the Machine Perception Lab, is a prototype system that classifies facial expressions into seven emotion dimensions: neutral, angry, happy, disgusted, fearful, sad, and surprised. The system will be integrated into interactive books in the near future.

6. Summary

Taken together, the Conversational Agent Toolkit, CU Animate system and computer vision technologies developed at the Machine Perception Lab provide a set of foundational tools and technologies that can be used to develop perceptive animated interfaces. Our goal is to teach children to read and learn better from text through great learning experiences. Thus far, we are enjoying the journey immensely. All of the tools and technologies developed during the project are either freely available now [7], or will be available to university researchers and educators for non-commercial use when they have well tested and documented.

7. Acknowledgments

8. References

[1] R. Cole et al. Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction. IEEE Proceedings, Special Issue on Multimodal Systems. In Press.

[2] J. Gratch, J. Rickel, E. André, N. Badler, J. Cassell, and E. Petajan, "Creating interactive virtual humans: Some assembly required," *IEEE Intelligent Systems* vol. 17, no. 4, pp. 54-63, July/August 2002.

[3] B. Pellom, "Sonic: The University of Colorado continuous speech recognizer," Center for Spoken Language Research, University of Colorado, Boulder, CO Rep. TR-CSLR-2001-01, 2001.

[4] B. Pellom and K. Hacioglu, "Recent improvements in the Sonic ASR system for noisy speech: The SPINE task," in *Proc. IEEE Intl Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[5] W. Ward, "Extracting information from spontaneous speech", in *Proc. of the Intl Conf. on Spoken Language Processing*, Tokyo, Japan, 1994.

[6] J. Ma, J. Yan, and R. Cole, "CU Animate: Tools for enabling conversations with animated characters," in *Proc. Intl Conf. on Spoken Language Processing*, Denver, CO, 2002.

[7] CU Communicator download (2002). [Online]. Available: http://communicator.colorado.edu