Towards a Roadmap for Speech Technology

Thursday, September 4 2003

Special session at Eurospeech 2003 in Geneva (September 1 - 4)

Organized by ELSNET

Report by J.M.Pardo

The session was scheduled as follows: An introduction to the Roadmap concept and the objective of the session by Steven Krauwer followed by four invited papers presented by Paul Heisterkamp, Bjorn Granstrom, Ron Cole and Roger Moore. The session ended with a discussion on the topics presented and on the general roadmap exercise with questions from the attendants. The schedule of the session is presented below:

13:30-13:35	Introduction	Steven Krauwer	ELSNET/Utrecht University
13:35-13:55	"Do not attempt to light with match!": Some thoughts on progress and research goals in Spoken Dialog Systems	Paul Heisterkamp	DaimlerChrysler
13:55-14:15	Multimodality and speech technology: Verbal and non-verbal communication in talking agents	Björn Granström and David House	KTH (Royal Institute of Technology)
14:15-14:35	Roadmaps, Journeys and Destinations Speculations on the Future of Speech Technology Research	Ron Cole	Center for Spoken Language Research
14:35:14:55	Spoken Language Output: Realising the Vision	Roger Moore	20/20 Speech LTD
14:55-15:30	Panel and Discussion	Steven Krauwer and José Manuel Pardo	ELSNET/Utrecht University and Universidad Politécnica de Madrid

Comments to the papers presented in the Special Session

1) The paper by **Heisterkamp** addressed some of the problems and solutions that we will encounter today in Spoken Dialogue Systems and it mentions how we should teach people to use the systems. Main actual problems are due to people speaking not to the system, people not saying what they want, people not providing the information requested by the system. Heisterkamp gives some examples where semantic comprehension would be very difficult to achieve for a long time to come by technical

systems, many more than ten years. People don´t always mean what they say neither they say what they mean.

It also covers the fact that many systems worked today successfully through conventions and not necessarily through logical and natural behaviour. For instance the qwerty layout of the typewriter was NOT ONLY designed to write faster, but to avoid mechanic conflicts between consecutive keys. Today it is a convention and everlybody uses it.

One of the conclusions is that it would be good to establish Spoken Dialogue System (SDS) conventions instead of trying to match with a machine exactly a human behaviour. The final ultimate goal of a SDS system should be the one of a good, better, cheaper, convenient an reliable service instead of matching the human process. Naturalness and ease of use are not necessarily the same and to know how to use a system we need conventions and training, not necessarily a natural system since this definition is inherently impossible to establish. It also address the important point of making investment in designing a good dialogue, a topic not always well taken into consideration that make the systems fail.

2. The paper by **Granstrom** covered some of the problems related to the use of Multimodality today (i.e. integrating of audio and visual modalities) and how to solve them. The paper is advanced because it is discovering some of the models of using multiple signals and integrating them in a complete communication process. By its own nature, the paper although advanced, presents the state of the art of actual systems (it does not attempt to predict the future). Nominally it presents three problems, how to obtain data, how to model them and how to exploit it in dialogue systems. In the presentation, some demonstration of facial synthesis was done, emphasizing the holistic nature of the speech communication process. Three applications of facial synthesis were presented.

3. The paper by **Cole** sets his concept of Roadmap. First the objective has to be set up, and next the kind of journey we want to make to arrive to the objective. It correctly, in my opinion sets the objective: to achieve Great communication. and the kind of journey : characterized by many explorations, and guided by successes and failures during these adventures. Guided by ambitious goals and conducted by independent researchers. Under this point of wiew it would be impossible to establish some predictions because they depend very much on the successes and failures of the researchers, between other parameters. The paper also points out the parameters that define Great Communication. Emotional, immersive ad personal. It hypothesizes also that the evaluation of future systems, taking into account this view would be more related to the usefulness or not of the experience of the users using the systems.
The papers remind us about the multidisciplinarity of the problem to attain the objective: Speech Research, Psychology and Cognitive Sciences, Linguistics, Computer Science, Electrical Engineering. It is important to establish a good interdisciplinary team with experts in all these disciplines.
The paper shows some steps that the Colorado team are doing in this direction. The opinion of Cole clearly proposes future systems closer and closer to human behaviour.

4. The paper by **Moore** is dedicated to a particular Roadmap exercise on Spoken Language Output. It has the ingredients that we are looking for: we want to know what will happen in the mid-term future in the area and the possible steps needed to make it happen.

In contrast with Cole´s vision, the Roadmap is defined with the objective (where to go?) and the optimum way to achieve the objective (how to get there?) and not the nature of the trip (what kind of journey we want to get there?).
It is also driven by Market pull, trying to match it to Technology push. This view is much more practical and realistic of what will possibly happen and it matches also any reasonable plan for an industry involved in the field. First the market opportunities are identified and them the product feature concepts that could satisfy them are defined and finally the technical solutions required to realise the new products.

The Market drivers are identified from the 6th EU framework programme: "a future in which computers and networks will be integrated into the everyday environment, rendering accessible a multitude of services and applications through easy-to use human interfaces" Although not labelled with time marks, some technical challenges are listed in the Spoken Language Output task: Improved modelling of style ,voice, and prosody, better modelling of the vocal tract and –very interesting- models that learn, models with proprioceptive feedback that hear and monitor their own performance.

Other papers presented at Eurospeech relevant to the Roadmap exercise

At least, three other papers presented at Eurospeech 2003 are relevant to the Roadmap exercise:

Speech and Language Processing: Where Have We Been and Where Are We Going?–
(Kenneth Ward Church)

This paper speculates with the future and with several questions:

a) More data is better data? The progress of Language processing has been alternating from data models to knowledge models. The start was data models (20 years) then there was a move towards knowledge models (grammars, rules) to constrain data models. We are actually in an era of data models again (we are in the second decade of it). The prediction is that in 10 years time we WILL HAVE to go back to knowledge-based models.
b) What will we do with petabytes of data that will be available (10 exp 15)? Search will become a key problem and a model related to how to do it will be important.

ISCA Special Session: Hot Topics in Speech Synthesis
Gerard Bailly, Nick Campbell, Bernd Möbius

What are the Hot Topics for speech synthesis? How will they differ in 5-years time? ISCA's Special Interest Group on Synthesis presents a few suggestions. This paper attempts to identify the top five hot topics, based not on an analysis of what is being

presented at current workshops and conferences, but rather on an analysis of what is NOT. It will be accompanied by results from a questionnaire polling SynSIG members' views and opinions. The fact that it abstracts the opinions of several recognized experts in the area makes it meaningful. It mentions evaluation, extension, emotion, multimodal and "type of input to the synthesizer" as key topics today that will still be alive in 2008.

A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners

Roger K. Moore

Again we have here the contribution of Roger Moore on an important topic, this time speculating with the speech recognition task.
Since the introduction of hidden Markov modelling there has been an increasing emphasis on data-driven approaches to automatic speech recognition. This derives from the fact that systems trained on substantial corpora readily outperform those that rely on more phonetic or linguistic priors.
Similarly, extra training data almost always results in a reduction in word error rate - *"there's no data like more data"*.
How much speech a human listen in a full life? : 120.000 hours. How much speech would be needed with actual system performance (extrapolating) to achieve human performance (70 life-times). The conclusion is: Our model of speech recognition training is much poorer than human´s. So more work would be needed to do in language models. It also mentions that with the same amount of speech, an automatic training system does better than a child in speech transcription.

Discussion

Some of the discussion was related to the presentations on Multimodal animated agents. One question arosen from the audience is that if the goal is to strive for naturalness of the agent, believability or the we rather want some system that can help us in any way, even if it is not similar to a human being: The answer is that the society will assess what application would be possible. We don´t know now. Another comment on the topic is that the animated agent will CHANGE the way a human person speaks to it, so they could finally sustain and effective communication.

Roger Moore warns us about the term naturalness. Naturalness is good but naturalness is a serious problem. He warns not to use this term. The conclusion again is that , what is considered natural? The human beings have been evolving during the years and today is natural something than yesterday was not.

Ron Cole concludes on his vision of the roadmap: Take big challenges and solve them