WORKSHOP

Past and future of speech technology, spoken language processing and intelligent multimedia

The guest speakers are:

Professor William J. Barry, Institut für Phonetik und Phonologie, Universität des Saarlandes, Germany
Professor Isabel Trancoso, INESC - Instituto de Engenharia de Sistemas e Computadores, Lisbon, Portugal
Professor John H L Hansen, Center for Spoken Language Research, University of Colorado at Boulder, Colorado, USA
Professor Paul Mc Kevitt, School of Computing & Mathematical Sciences, Faculty of Informatics
University of Ulster, Londonderry, Northern Ireland.

The schedule of the workshop is:

09.00 - 09.10 Opening - Paul Dalsgaard

09.10 - 09.55 **Phonetics for Speech Technology or Speech Technology for Phonetics? -** Bill Barry

Summary: Speech Technology and Phonetics are necessarily linked in their common object of interest – spoken language. For most speech technologists, however, Phonetics is a science which tries to answer questions that appear to have no real relevance for Speech Technology, and for many phoneticians, Speech Technology research appears to have nothing at all to do with understanding the processes involved in speech production and perception. Against such a background of quasi-separate cultures, where it appears that progress in either area is achieved on separate paths through the wilderness, the questions need to be asked, whether the paths are converging or diverging (or merely proceeding in parallel) and whether they have been more closely linked at times in the past. In both scientific sub-communities we need to consider whether mutual understanding can contribute to progress for both groups in the goals as defined at present (or for neither group). We also need to reflect on the changes to self-understanding that contacts inevitably bring, and how this affects the scientific paradigm we work in.

09.55 - 10.40 **The Alert system for selective dissemination of multimedia information -** Isabel Trancoso

Summary: The goal of the recently finished ALERT European project was to build a system capable of continuously monitoring a TV channel, and searching inside their news programs for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a broadcast news program. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send to that user indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the video clips referring to the selected stories. The last stage of this project was the field trial of the alert prototype.
The Portuguese prototype significantly differs from the ones developed by the other

participating countries, because Radio Televisao Portuguesa (RTP), the Portuguese user partner in the ALERT project, is interested in indexing every story and not only the stories according to certain profiles. To accomplish this indexing task, we based our topic concept in a thematic thesaurus definition that was used at RTP in their manual daily indexing process. This thesaurus follows rules which are generally adopted within EBU (European Broadcast Union) and has an hierarchical structure with 9 levels and 22 thematic areas in the first level. In this presentation, we shall describe the three main blocks of the system: the CAPTURE block, responsible for the capture of each of the programs defined to be monitored, the PROCESSING block, responsible for the generation of all the relevant markup information for each program, and the SERVICE block, responsible for the user and database management interface. We shall particularly emphasize the main stages of the PROCESSING block. A simple scheme of semaphores is used to control the overall process. The system has been implemented on a network of 3 machines. These are ordinary PCs running Windows 2000 and Linux. We shall end by describing the recent field trials and presenting a demo.

10.40 - 10.55 Coffee break

10.55 - 11.40 **Spoken Document Retrieval for a National Gallery of the Spoken Word: `One Small Step'** - John H.L. Hansen

Summary: In this talk, we consider emerging challenges in the development of an online spoken document retrieval system, SpeechFind, for a National Gallery of the Spoken Word. As part of an on-going U.S. NSF Digital Library Initiative II project, SpeechFind is intended to serve as an audio index and search engine for the largest spoken word collection in the world spanning the 20th century with as much as 60,000 hours of audio archives (from T. Edison's first cylinder recordings, to famous speeches such as man's first steps on the moon "One Small Step", to American presidents over the past 100 years). This is a partnership between CSLR-CU and Michigan State University (MSU: ECE, Vincent Voice Library, Matrix). The group at CSLR-CU is responsible for the audio search and text indexing engine, while researchers at MSU are responsible for digital watermarking, digitization, metadata construction, and educational tool development. In this talk, we will consider a number of challenges in developing automatic speech recognition technology to address the wide range of recording conditions seen in automatic transcription. We will also consider issues relating to text based search issues.

11.40 - 12.25 **Intelligent MultiMedia and StoryTelling** - Paul Mc Kevitt

Summary: There is a need for semantic representations that can bridge the gap between linguistic inputs and their corresponding visual knowledge which are indispensable in performing a variety of tasks involving the automatic generation of 3D animation. The semantic representation of events in visual knowledge and the design of a suitable knowledge base specifically for the integration of linguistic and visual information are discussed here. We describe a framework used to represent action verb semantics in a visual knowledge base. Visually observed events are described by establishing a correspondence between verbs and the visual depictions they evoke. The method proposed here is well-suited to practical applications such as automatic language visualisation applications and intelligent storytelling systems. In particular, it will be useful within CONFUCIUS, a system which receives input natural language stories and presents them with 3D animation, speech, and non-speech audio.

12.25 - 12.45 What’s coming next? - Paul Dalsgaard

13.00 - Light Lunch buffet in the NOVI canteen.