i M i
INTERACTIVE
MULTIMODAL
INFORMATION
MANAGEMENT

The (IM)2 Newsletter

Every month the (IM)2 Newsletter brings you the latest and hottest
scientific and administrative news about the (IM)2 NCCR and related topics

News from the (IM)2- ICSI exchange agreement

(IM)2 and the International Computer Science Institute in Berkeley have an exchange agreement that allows young researchers to spend some time in the other institution. We talked to the first two swiss students to visit ICSI, Jitendra Ajmera from IDIAP and Micha Hersch from EPFL:

Jitendra and Micha at ICSI.

What were you doing (and where) before you went to ICSI ?

JA: I was working towards my PhD (second year) at IDIAP before I came to ICSI. My PhD work revolves around automatic segmentation and clustering of large multimedia documents.

MH: I was a computer science student at EPFL, Lausanne.

Where did you learn about the IM2-ICSI exchange program ?

JA: I learnt about this program in IM2 kick-off meeting held at EPFL sometime in March last year.

MH: Through Herve, whose class I attended at EPFL

Did you know ICSI before ?

JA: Yes. I learnt about ICSI as soon as I joined IDIAP through some anecdotes from my supervisor about his work at ICSI. I also knew some people here from literature surveys in journals or on the internet.

MH: I had heard of it through Herve

What project are you working on at ICSI ?

JA: I am working on a project called EARS (Effective Affordable Reusable Speech-to-text).

MH: Automatic speaker and subword units clustering in speech processing. I am taking a clustering algorithm developped at IDIAP, optimizing it, and using it to develop a speech recognition system, which is not based on any linguistic knowledge,

as opposed to other standard speech recognition systems. This work constitutes my diploma thesis for EPFL.

How does your work contribute to ICSI and how does your work at ICSI fit into your original project ?

JA: ICSI and IDIAP participate in a lot of common projects and themes. Thus, I can apply some of my work directly to the projects that are being handled at ICSI. It helps me also because I can test and verify my previous conclusions.

My main focus presently is investigating the possibility of automatically extracting subword units from the audio data, which again is of common interest to both institutes. I am working on this with the help of my colleagues at ICSI.

I am benefitting from this in various ways. First, I am getting to interact with other people and their ideas, which I think is what a researcher always needs. Second, we are sharing some work between us so that we don't reinvent the wheel between two institutions. Third, I am learning about areas of research other than mine from various projects that are going on here. We are also holding a reading group between some of us at ICSI and I find it really helpful.

MH: Researchers at ICSI are very interested in automatic subword unit derivation, and new models for speech recognition. I am working in those two areas. They are also interested in automatic speaker clustering, and especially need it for mandarin speech recognition.

I have been given a lot of freedom in my work, which enables me to do exactly what I wanted to do.

Do you expect to keep in touch with the people you met at ICSI after you return to Switzerland ?

JA: Of course! I have met some excellent people here and I would certainly like to keep in touch with them after I return.

MH: Yes

What do you recommend to other people interested by the IM2-ICSI exchange program ?

JA: For IM2 people, I would say that it is an excellent opportunity. At the same time, if possible, I would also like to see some people coming to IDIAP from ICSI.

MH: It is good to come with a project that is of strong interest to some of the researchers at ICSI. This will make collaborations with them more easy and fruitful. Since one has a lot of freedom in the work, people at ICSI will not discourage someone with a project of no big interest to them. But this person might well end up working alone on this project, with little help from and interaction with other researchers.

Anything else you would like to add ?

MH: As I am just finishing my degree, this exchange is a wonderful and most interesting experience, as I discover the realm of scientific research. The San Fransisco Bay area and Berkeley in particular are also very interesting and dynamic places.

i M i
INTERACTIVE MULTIMODAL INFORMATION MANAGEMENT

ICSI

www.im2.ch/icsi-exchange.php

The (IM)2.SA Workshop

The first (IM)2.SA workshop was held at EPFL on February 5th 2003. It gathered 28 researchers from all the participating institutions (EPFL, ETHZ, IDIAP, UniGE, UniBE).

The program was split in two sections. In the first part, four core projects and five white papers have been briefly summarized by the respective project leaders. In the second and most important part, 17 Ph.D students involved in (IM)2.SA presented their work in detail. Very fruitful discussions revealed a large number horizontal topics and synergies. The program, the participants and their presentations can be found in the internal pages of the IM2 website. The next (IM)2.SA workshop is planed for February 2004.

Events

(IM)2 Scientific and Industrial Advisory Boards   13–14.2.03

The first joint meeting of the (IM)2 Scientific and Industrial Advisory Boards took place in Martigny on February 13 and 14. We will report back on the outcome of the meeting in the next edition of the Newsletter once we have the report and recommendations of the Boards.

iM
INTERACTIVE
MULTIMODAL
INFORMATION
MANAGEMENT

The Speech Processing Group at TIK/ETHZ

Institut für Technische Informatik und Kommunikationsnetze

The speech processing group (SPG) was founded in 1976 and started work in speech coding and noise cancellation. Later on, the scope was extended to further topics, namely text-to-speech synthesis (TTS), text-dependent speaker verification, and speech recognition.

In the nineties, we have focused our research activities on TTS synthesis and speech recognition. These two topics share a common peculiarity, namely their involvement in both surface structures of natural language: text and speech. The aim is to transform one surface structure into the other, i.e., text into speech or vice versa.

It is commonly acknowledged that this cannot be achieved without the linguistic knowledge of the language(s) in concern. In order to be able to recognize speech from a specific language, we must know the set of phonemes of this language, the words that belong to this language and their pronunciation, which words can form expressions and sentences, etc. There is no consensus among researchers about the representation and the application of such knowledge, however.

Our intention is to use always that type of representation which matches the type of knowledge to be represented. Thus, we consider phoneme characteristics to be adequately represented by statistical means (e.g. hidden Markov models), whereas sets of appropriate rules are meant to be suitable for the description of morphological and syntactic knowledge. This in turn leads inevitably to inhomogeneous systems, that include statistical parts and knowledge-based parts, and therefore are generally significantly more intricate to be realized.

We do not flinch from going that way, however, because it is considered to lead finally to the better solution of the problems than e.g. the purely statistical approaches. We are already able to demonstrate the correctness of this view in the context of polyglot TTS synthesis (see the figure) and we are confident to attain this point one day also for speech recognition, which is admittedly more challenging (more information: www.tik.ee.ethz.ch/~spr).

SPG is happy to contribute also to IM2, particularly with polyglot TTS synthesis. Such TTS will e.g. be an important component of speech-based user interfaces.

The SVOX AG spin-off

SVOX

While the SPG at TIK/ETHZ studies primarily the fundamental issues of speech processing, SVOX AG, an almost three year old spin-off of ETHZ, is concerned with development and marketing of TTS software products that make computers and other electronic devices talk by converting written text into natural-sounding and easily comprehensible speech. SVOX aims to be the acknowledged leader in the TTS market of embedded and server-based systems. SVOX AG has launched innovative TTS software components that can be easily integrated into speech applications by means of standard speech application interfaces.

The SVOX TTS is currently available in a wide range of language versions including German, French, English, Italian, and Spanish. With SVOX's software architecture customers are offered a TTS engine adaptable to their technical and market needs. While headquartered in Zurich, SVOX targets worldwide markets and runs representations in Germany, Austria and the US. See www.svox.com for more information about SVOX AG.

Permanent staff and PhD students of the SPG at TIK/ETHZ. From left to right, back: Michael Haslebacher, Philipp Frei, Harald Romsdorfer; front: Beat Pfister, Ulla Glavitsch and René Beutler.

bp

Mein Cousin surft im World Wide Web.

(a)

S_G
NG_G   V_G   PNG_G
POSS_G N_G   VST_G VE_G   PREPC_G   NG_G
mein   N_F   VST_E   t   im   NG_E
cousin surf   MOD_E MOD_E N_E
                          N_E   ADJ_E web
                          world wide

(b)

\G\main- \F\ku-[1]zɛ̃- #{2} \E\[2]sɜrft- \G\|im- \E\[3]wɜrld- [4]waɪd- [1]web.

(c)

mein   Cousin   surft   im   World   Wide   Web

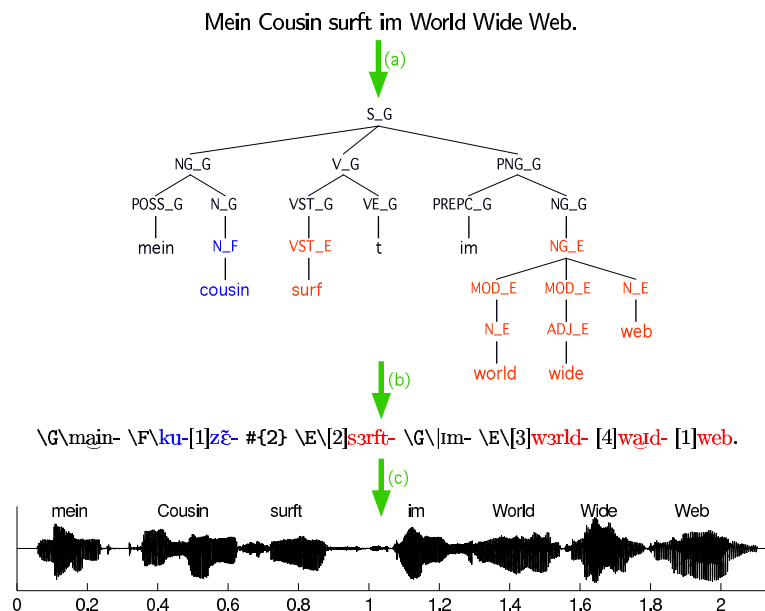0   0.2   0.4   0.6   0.8   1   1.2   1.4   1.6   1.8   2

Illustration of the major steps of our polyglot TTS synthesis: A mixed-lingual input sentence is at first processed by the morpho-syntactic analyzer (a). The output is a syntax tree displayed here in a simplified form. It shows among other things, that this German sentence (S_G) contains French (blue) and English (red) inclusions. Inclusions can range from a part of a word, e.g. the English verb stem "surf" (with German ending "t"), up to large constituents. The next processing step (b) assigns the syllable stress levels (in brackets) and sets the prosodic phrase boundaries (in curly brackets). The result of this linguistic processing is an abstract representation of the speech to be produced. From this representation the speech signal is generated (c) which includes mainly prosody control (i.e. determining fundamental frequency and duration of each phone by means of neural networks) and diphone concatenation.