Part V

Beyond CHIL

30

Beyond CHIL

Alex Waibel

Universität Karlsruhe (TH), Interactive Systems Labs, Fakultät für Informatik, Karlsruhe, Germany

Despite tremendous progress, the CHIL project represents a milestone rather than the final vision. Whilst many questions have been answered, the project has also opened up new challenges and directions for further exploration, and more work remains to be done.

Three main areas of concern have been explored in CHIL and are presented in this book: Perceptual Technologies, Services and Infrastructure. In CHIL, the project, of these, Perceptual Technologies has received the greatest attention primarily because when CHIL was initially proposed the perceptual technologies available at the outset of the program simply did not offer the robustness necessary to permit the development of flexible, perceptually informed CHIL Services. Nevertheless, the CHIL program managed to propose and showcase a set of initial CHIL services within its short lifetime and managed to evaluate and examine their effectiveness for real users. This was made possible via architectural and organizational tools and processes that were designed explicitly for the purpose of rapid prototyping and exploration.

At the conclusion of the CHIL program, it is clear that research will continue on all fronts, perceptual components, infrastructure services, and that further advances in perceptual technologies will also lead to more advanced and more daring new CHIL services in the future. The following comments speculate on future directions for CHIL computing, based on ideas and insights learned in the course of the three year CHIL effort.

30.1 Perceptual Technologies

Among the areas of concern in the CHIL program, Perceptual Technologies has received the most attention, and tremendous improvements have resulted from the effort, thanks in part to concerted, worldwide benchmarking efforts carried out in each of the perceptual processing technologies considered. The benchmarking has proven to be extremely beneficial for community building and to achieve a focused intense effort around common databases, leading to rapid progress, delivering high quality results and greater robustness in a short time frame. Robustness has benefited from

the concerted technical effort in each of the processing technologies as well as innovative work on fusing them across modalities.

- Robustness in Individual components: CHIL broke new ground in building all its benchmarks around real data collected in real environment, with real people carrying out real tasks. The program did – on purpose – not resort to artificial scenarios, intrusive sensors, or slow expensive processing technology. Everything had to work in real-time, and all the data came from real meeting and seminar rooms in real organization. The data was also collected at multiple (5) sites to ensure generalization of the results across idiosyncrasies of each local environment. Yet, numerous challenges still exist:
  - Environments: Despite the generalization across spaces explored in CHIL, all of our environments were still meeting rooms, and did not cover the full range of human spaces, such as corridors, lobbies, airports, railway stations, shops, offices, restaurants, streets, the outdoors, and many more. To generalize perceptual processing to all human experience, new and different environments and transitions between them need to be included.
  - Sensor Positioning: At the conclusion of CHIL good speech recognition and speaker identification accuracies have been achieved over lecture and meeting data, with close speaking or lapel microphones. Significant advances were also observed with remote microphones and microphone arrays. Nevertheless, even though error rates have dropped, sometimes from 70% to 20%-30%, the remaining errors are still too high for certain applications, where error rates below 10% are necessary. Further research will be required.
  - Interferences: In both acoustic and visual sensing, different environments bring greater interference. In speech, the most well known is the so-called cocktail party phenomenon, the apparent ability of humans to follow a conversation at a cocktail party, despite an overwhelming level of jamming noises and jamming human conversations, and despite the distant and variable positioning of a listener's sensors. Toward the end of CHIL, such questions were raised and corresponding challenges formulated, but the problem remains largely open and unaddressed. In vision, analogous interferences exist. Here too, our environment (meetings and lectures) provided a real but comparatively benign version of these. Railway stations and large crowded places still present greater challenges in resolution, noise and occlusion. Moreover, they may require a combination of different techniques aimed at dealing with long range vs. short range perception, and with the smooth integration of entirely different techniques available at high resolution vs. low resolution (for example, in the case of person identification: face ID vs. gait or color histograms of clothing).
  - Sensor Coordination: A surprising discovery in CHIL was the challenges and opportunities emerging from the coordination of multiple sensors. When perceptual processing was done only in directed, well positioned human-machine tasks, there was typically one well positioned sensor from which data was collected such as a camera straight ahead or a close-speaking, head-

mounted microphone. Similarly, the phases of recording were well defined, by way of an on/off switch, or a shutter release at the right moment. CHIL removed this artificial constraint, and instead placed multiple sensors in a space. With multiple cameras observing the same scene, and multiple microphones listening to the same acoustic events, the question of coordination and integration became paramount. In CHIL, several calibration techniques were successfully tried and the signals collected from multiple sensors combined for the benefit of several of the perceptual processors (e.g., speech recognition, focus of attention tracking, people tracking, etc.). Despite these advances, more is clearly going to be required such as how do sensors know where they are in any environment and how do they communicate and merge their results more effectively? How do sensors know when the collected signal is unreliable how does the perceptual processor identify the most informative signals from multiple sensors and varying reliability?

- Robustness through Fusion: CHIL examined a number of perceptual tasks, where multiple modalities cooperate to describe human communicative events. Speaker localization, or person identification, for example can be done based on the acoustic signal as well as the visual signal. A number of results from such multimodal processing have been reported in this book, and will continue to attract research interest moving forward. In addition to work on the mere combination of multimodal signals, two new research directions have emerged:
  - Opportunistic Multimodal Fusion: With multiple signal streams from different modalities and multiple sensors, the same event cannot always be detected at the same time and may vary in robustness and reliability. A speaker may, for example, speak or be silent, a face may be temporarily occluded and different cameras or microphones may yield better and more reliable signals at different times. Fusion must therefore be selective and accumulate evidence over time. Identifying such moments of high robustness, better confidence measures and better integration across time, will continue to drive research as we aim for increasingly natural interactions and environments.
  - Self-Calibration: In addition to opportunistic fusion, we must also consider a more adaptive approach to achieving perceptual robustness. For a start, sensors must be able to self-calibrate better, and identify their own positioning and their own role in performing a perceptual task, vis a vis the other sensors. This is particularly important if we hope to build flexible, and general, perceptual components for practical commercially relevant deployments. For such deployments, the sensors and their precise positioning will vary and the arrangement cannot be redesigned on site in a cost-effective manner: The sensors must therefore arrange and determine their cooperation by themselves.
  - Active Fusion: Beyond Opportunistic Fusion and Self-Calibration across fixed sensors, we may also consider the possibility of sensors that perform more active perceptual processing by moving into position. This is of particular interest in applications that offer the possibility of moving the sensors during processing, such as humanoid robots, vehicles, or transportation systems. A moving platform could thus be positioned to take a "better look" or

turn to better listen in from different angles and distances. Such multimodal perceptual processing will require considerably more complex models of the perceptual systems and their environments.

30.2 CHIL, A Family of Services

CHIL computing is not and was never meant to be limited to the four CHIL services, considered in this book. Rather it represents a vision for numerous proactive services that aim to support human interaction without necessitating direct commands or human-machine interaction. Numerous additional services are possible. Even during the project, several such "surprise" services emerged at the participating laboratories: The "hummer", a system observing and modeling human conversational speech and turn-taking, the "Lecture Translator" a simultaneous speech translation system for seminar speakers with selective audio presentation for select subgroups in an audience, or a meeting coaching system for consultants, are among the surprising outcomes of some of the work in CHIL that has already occurred during the life-time of the project. Further advances are likely, and are facilitated by the CHIL architecture and the availability of interchangeable perceptual modules among the partners. Security applications, advertising, coaching, and assistance to the elderly, are further applications that appear now possible and that are already being considered beyond CHIL. And architecture for assembling components into CHIL services, the methodology for evaluating perceptual components and evaluating usability of the resulting systems permit a rapid prototyping to explore CHIL services beyond the ones discussed in this book. Indeed, we very much hope that we have only just scratched the surface.

30.3 From CHIL to CHHIL Services

In considering CHIL systems, we began with the rather extreme position of exploring the "disappearing" computer and the technologies that could be necessary to make this reality. We felt that such an extreme position was necessary to drive progress and to lay the groundwork for implicit computing capabilities. Robust perceptual technologies, implicit computing services, and flexible architectures, have all been advanced and benefited considerably from taking this view, as they are key elements of truly flexible computer systems in natural human environments.

Nevertheless, for many practical systems, there is no need to make a hard decision between implicit CHIL systems and services and more traditional human-machine interaction and dialog. In fact, it is useful to think of human-machine interaction as part of CHIL services or of Computers and Humans in the Human Interaction Loop (CHHIL). Already in CHIL, we have carried out work on human-machine dialog, suggesting a reappearing computer that not only observes but also listens and occasionally "comments" on the interactions and communications between humans. Indeed, a truly useful autonomous device of the future may take a more balanced

approach between pure CHIL and Human-Machine interaction and seek a middle ground between implicit and explicit interaction. Why should a perceptually well-informed, proactive, and intelligent social agent not engage in an occasional direct dialog with its human partner, while also at the same time proactively taking the initiative? There are numerous scenarios that would make this an attractive proposition:

- Humanoid robots that interact occasionally with their masters, but that are capable of doing their work autonomously
- Smart rooms with avatars occasionally speaking up, or engaging in an occasional directed dialog
- CHIL services, that are occasionally addressed explicitly by humans in the room
- Learning CHIL services that occasionally request clarification or instructions from humans

All this, will necessitate further exploration of study of the interplay and trade-offs between direct/explicit and indirect/implicit interaction, and between autonomy and explicit command & control. Attaining a balance between these will indeed be challenging. Not only will it involve further advances in CHIL Computing, Human-Machine Interaction, and their technical integration, but also raise new social and philosophical issues. How is a balanced integration achieved? When should a computer system speak up, when to interrupt and when to patiently observe? What social norms are to be applied and how? How does the system assess relevance and urgency in interacting with or in the presence of humans? How does a system decide whether to take on the initiative and proceed proactively and when to await instructions? And finally, how much autonomy would we want and how much would we be willing to yield to a computer artefact and under what circumstances and tasks?

For the moment, these questions are only suggestive for potentially profitable ongoing research, whilst work on more mundane but no less challenging tasks, remains. Despite the impressive advances so far, it is clear that further work is needed to advance our understanding of perception, cognition and human interaction in order to achieve computer systems that will blend in with humans and as gracefully as humans, in a community and a social partnership.