Graphical user interfaces (GUIs) have been around for more than a student lifetime and have proved their worth in thousands of applications that are being used every day. From an ergonomic point of view, GUIs are little more than a glorified collection of the knobs and dials familiar from ante-computer technology: you manipulate with your hands, watch what happens, do action repair if needed, and go to the next step. If you are interested in computers with human capabilities, vision and speech open an entirely new world of computers that can *see* and *talk* like we do. Computer vision is the moody input cousin of computer graphics – you have all the time you can afford to program the rendering but visual input is unpredictable and messy reality. Computer speech is both input and output, like in systems capable of spoken dialogue. Once we have computer speech and vision both, the system and we can see the same objects and events *and* talk together about them, and the system can capture the rest of our communicative behaviour as well – facial expression, gaze, gesture, body posture, walk, object manipulation. Researchers across the world are beginning to address the enormous application potential of this multimodal scenario, and speech and vision people are getting together like never before.

Arguably, viewed as enabling technologies, computer speech still holds the maturity lead over computer vision. Even though the speech signal is enormously rich in information and we are still far from mastering important aspects of it, such as recognition and on-line generation of the speech prosody which people use voluntarily for emphasis or semantic disambiguation, and more or less involuntarily for many other things, like emotion and physical state expression – it is still much easier to shut up the people in a room to get a clear speech signal than to control its lighting conditions and identify and track all of its 3D contents independently of the viewing angle. Still, EyeToy is already out there and you even control the game "by making some noise". Once that noise gets replaced by speech, another challenge appears, one which was first described in Bolt's 1980-"Put-that-there" paper. This is the problem of combining the semantics of input in different acoustic and visual modalities, such as speech and pointing gesture, into a single coherent message through what is called semantic-level multimodal fusion. At this point, we don't even have a satisfactory conceptual framework for addressing this problem in its general form. There is also signal-level multimodal fusion of speech and vision, for instance in audio-visual speech recognition which combines speech signal information with mouth and lip movement information to improve speech recognition.

Thus, given the state of the art, it makes good sense that the papers in this issue of Crossroads are about speech *or* vision. Three articles address different stages of the process of making computers understand what is commonly called the speaker's communicative intention, i.e., what the speaker really wishes to say by uttering a sequence of words. Deepti Singh and Frank Boland discuss approaches to the important pre- (speech)-recognition problem of detecting if and when the acoustic signal includes speech in the first place. If no speech is present, there is no reason to spend computational resources on recognition or speaker identification, nor, perhaps, to steer a camera towards the source. Focusing on automatic speech generation, or synthesis, Claire Brierley and Eric Atwell provide a vivid illustration of the continuing debate over the extent to which natural language or, indeed, human communication more generally, is rule-based or stochastic. While a couple of simple rules appear to capture large fractions of the cases in which people pause during speech in order to chunk their speech for semantic intelligibility, it is far less clear if it is possible to capture the remainder by adding more rules. Nitin Madnani's introduction to natural language processing, or NLP, is likely to tempt computer scientists to try out NLP for themselves. NLP is not only essential to the development of the language models which make speech recognisers recognise well, it is also key to extracting the semantics of the spoken message.

The two articles on machine vision make two equally interesting points about the present state of the field. Taking human faces as an example, Justin Solomon compares the relative ease with which

it is possible to solve complex face rendering problems with the difficulty of modelling the unique face each one of us has. Gang Gao and Paul Cockshott describe how smart use of computer image processing promises a robust shortcut solution to the integration of MR images of the same object generated using two different imaging techniques.

Niels Ole Bernsen 29.1.07