The squid that hid

or camouflage as a (mis)understanding of context

In computing, *speech recognition* is the translation of spoken words into text, and its performance is measured in terms of accuracy and speed. Speech recognition by a machine is a very complex problem. Human vocalisations vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume and speed, all of which may be distorted by background noise, echoes and interference. And perhaps the most difficult obstacle of all, language as it is naturally spoken doesn't contain breaks between words. Instead, the words blend together, making it very hard for a computer to tell where one ends and another begins.

A squid is an elongated, fast-swimming cephalopod mollusc with eight arms and two long tentacles, typically able to change colour. The word *squid* is of uncertain origin but is thought to be a sailor's variant of *squirt*, so called for the ink it squirts to baffle its predator and escape from danger. The 'sounds like' of this etymology is echoed in the 'looks like' of squid camouflage. Using a combination of chromatophores (tiny muscle-controlled bags of pigment in the skin) and iridophores (cells which can reflect different wavelengths of light, i.e. different colours) the squid is almost instantaneously able to control its transparency or match its background perfectly and hide. The problem of how squid are able to choose particular skin colours to camouflage themselves so successfully is particularly interesting as their eyes are completely colourblind. Recent research has found that squid skin contains light-sensitive proteins called opsin, leading to the conjecture that the squid's skin may check the environment itself, cell by cell - not via the eye or brain - to see what colour it should become. In an act of total understanding of context, the squid weaves itself into its surroundings with speed and accuracy.

Computer speech recognition essentially seeks to translate information from one state to another - from speech to text. To do so, a whole chain of material manipulations and complex transformations have to take place. First, the spoken words - vibrations in the air - are captured and converted to a digital signal by taking precise measurements of the wave at frequent intervals. The digitised sound is filtered to remove unwanted noise and sometimes to separate it into different bands of frequency (what we hear as difference in pitch). The sound is then normalised to a constant volume and the speed adjusted through a process called 'dynamic time warp' to match the speed of the samples stored in the system's memory. Then the signal is divided into small samples - 100ths or 1000ths of a second.

Next and most spectacularly, the programme examines the samples in the context of the other samples around them. Most current speech recognition programmes use statistical modelling systems: hidden Markov models and neural networks. These models take information known to the system (the tiny, chopped up, digitised sounds) to figure out the information hidden from it (the sequence of words that have been spoken). In such models, all sentences in a language are permissible but some are more probable than others. By working out the probability ranking of different possibilities the likeliest sequence can be found. Probabilities of one section of a sequence can affect another, both forward and backwards, in a context-based system that is constantly building on, and creating, its own context. No speech recognition system achieves 100% accuracy, and accuracy diminishes as vocabulary size - potential context - increases. If the model 'misunderstands' the real context, the original message swims camouflaged in a sea of sounds-like. That is - insight is quick / inside the squid.

Anna Barham, 2014