At just a few months infants learn to distinguish simple speech patterns and categorise them according to probability. It is particularly easy for children to learn new words if the voice is familiar to them.

PHOTO: CORBIS – STOCKMARKET

also need this learning phase. To make a machine recognise first sounds, then whole words and sentences and finally long stretches of speech, scientists have to design algorithms – complicated series of logical or mathematical operations. For this, the speech signal is digitised and converted into a form which a computer can process. The machine has to create a reference pattern from each sound, it has to form acoustic units. At a subsequent stage, strings of such acoustic units are converted into words. To put it simply, the computer attempts to match existing frequency diagrams with the acoustic signals it has just "heard".

Practise makes perfect – this is also true of speech recognisers. The better trained a recognition system is, the fewer mistakes are made. Many of the programmes which have appeared on the market in the past are based on a relatively limited vocabulary. They can only understand quite specific types of information – such as simple messages. "If you want to place an order, please say 'one', if you want to speak with one of our advisors, please say 'nine'". This is the pattern followed by the type of speech recognition programmes currently popular, which could be described as command receiver. Automatic dictating machines have a much larger vocabulary, although they do have one distinct disadvantage: they are almost always adjusted to one individual speaker and are extremely sensitive to sudden changes in voice. The machine has to undergo a lengthy training period to understand unknown speakers. "Automatic speech recognition is based on statistics", says Roel Smits, "the more data, the more accurate the understanding."

Humans are obviously in a better position to recognise linguistic nuances in everyday conversation and to avoid phonetic traps. The English language is full of ambiguity. Homophones – words which sound the same – are common, presenting considerable problems for foreign speakers. To understand the difference between the pronoun "where" and the verb "wear", the computer needs to grasp the context. Similar difficulties are caused by pairs such as "threw"/"through", "son"/"sun" and "fare"/"fair". These often bother even native speakers but cause real problems for speech recognition programmes.

However, there are areas where machines are already superior to humans. When it comes simply to identifying speakers, computers can decode certain signals with great precision. There is virtually no risk of them being caught unawares by voice imitations: the frequency diagrams speak a clear language. Small wonder that automatic speech recognition, and automatic speaker recognition in particular, is playing an increasingly important part in criminology.

THEORY OUTSTRIPS PRACTICE

Yet, when it comes to recognising complicated speech units, there is room for considerable improvement in recognisers. Anne Cutler of the Max Planck Institute and Roger Moore of the English firm 20/20 Speech calculated, using models, how long it would take a machine to attain a virtually zero percent error rate: in theory, between two and nine million hours of training would be needed for a computer to reach the capacity of a near-perfect human listener. Perfection is not yet a term which can be applied to the current generation of computer programmes. However, companies such as IBM with its ViaVoice system are working flat out to make speech recognition more reliable. Less than one percent of the population currently use automatic speech recognition – yet that is all set to change soon.

The psycholinguist Roel Smits believes that, within a few decades, microchips in domestic appliances will recognise human commands and pass them on to intelligent machines. Remote control devices will also respond to verbal commands in the same way as PCs. Mobile phones already contain chips with speech recognition technology. Major advances in automatic hearing aids are also expected.

As is so often the case, theory far outstrips practice and this is especially true of speech recognition. Hynek Hermansky, who lectures at the Oregon Health and Sciences University in Portland, presented a recognition system at Nijmegen based on the structure of the human ear. This system uses intervals of one second of speech which contain up to 15 speech sounds. Differentiated speech recognition such as this would be far less sensitive to background noise or distortion than previous systems, which assess amounts of energy in intervals the length of a single speech sound or less.

So when will we be able to communicate with speech recognisers just like Captain Kirk in the Starship Enterprise? The workshop in Nijmegen was less concerned with concrete applications and with visionary projects which touch upon the field of artificial intelligence. The scientists were more interested in representing the state of basic research. For Roel Smits believes that, without a transfer of knowledge between human and automatic speech recognition, the new technology will not advance significantly. "We're up against a limit – even if we manage to increase computers' data capacity further." CHRISTIAN MAYER

Where do spoken words come from?

Core operations in normal speech production are the accessing of words in memory that appropriately express the intended message, and the preparation of each word retrieved for articulation. The theory developed in the MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS, *Nijmegen, The Netherlands, provides a detailed account of both mechanisms (PNAS, 98, 23, November 06, 2001).*

Every normal person learns to speak, and speaking involves, among other things, producing words. By reaching adulthood a speaker in our Western culture may well have produced some 50 million words. There is hardly any other human skill that is so well practiced. In normal speech we produce words at rates of some 2 to 4 per second. These words are continuously selected from a mental lexicon containing tens of thousands of words. Still, we make few errors. On average, we select the wrong word (for instance left when we mean right) no more than once in a thousand items. How is this robust, high-speed mechanism organized?

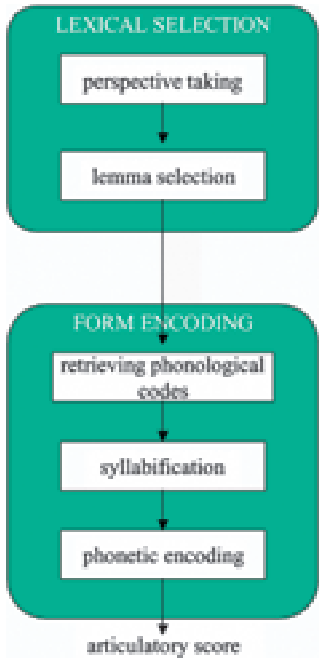
The theory proposed consists of two major processing components (Fig. 1). The first component deals with lexical selection. It is the mechanism that, given semantic input (some state of affairs to be expressed), selects one appropriate lexical item from the mental lexicon. The second component deals with form encoding. It computes the articulatory gestures needed for the articulation of the selected item. The theory has been computationally implemented under the name of WEAVER ++, and its experimental verification involved a decade of teamwork by Levelt's research unit at the Max Planck Institute, particular involving Drs. Antje Meyer and Ardi Roelofs.

A major experimental paradigm used has been picture naming. A picture to be named, for instance one of a horse, is presented to a subject. The instruction is to name the picture as fast as possible. We measure the latency from picture onset to the onset of articulation. This latency is about 600 milliseconds for naming a horse. Apparently, lexical selection and form encoding can be completed within two-thirds of a second. During lexical selection two successive operations are run. The first one, perspective taking, consists of selecting the target concept for expression. Experimental conditions can be manipulated such that subjects will use horse or animal or stallion to refer to the object. The second one, lemma selection, consists of selecting the one corresponding item from the mental lexicon, for instance the item 'horse'. The item is called a 'lemma', which roughly means 'syntactic word', i.e. the word's syntactic properties, such as word class (noun, verb, etc.) or other syntactic features (such as gender for nouns, or transitivity for verbs). Selecting the item appropriate to the target concept takes place under competition. Semantically related items, such as 'animal' or 'stallion', are measurably coactivated. The quantitative computational theory predicts the selection latencies for selection under competition. The theory is tested by way of picture naming experiments where subjects are presented with an auditory or visual distracter word while they name the picture. The distracter is to be ignored. If horse is the target name, hearing the unrelated word chair slows down the response latency by some quantity. But hearing the semantically related word 'cow' has an even stronger inhibiting effect, an extra 50-100 milliseconds (dependent on conditions). This so-called semantic inhibition effect has been tested and quantitatively confirmed in a large variety of experiments. The time course of lemma selection, predicted by the theory, was further tested and confirmed by way of magnetic encephalography (MEG) in joint work with the Max Planck Institute for Cognitive Neuroscience in Leipzig. That work showed, in addition, the involvement of regions in the left lateral temporal lobe in the operation of lemma selection.

Form encoding is initiated upon selection of the target lemma. The first step here is the retrieval of the target item's phonological code, an abstract string of phonological segments, for instance (h, o, r, s). Retrieving a word's phonological code is faster for words that are frequently used than for low-frequency words (by some 40 milliseconds). In picture naming, retrieving the code can be facilitated by providing the subject with a phonologically related distracter word. Subjects are faster in naming a horse if presented with a distracter such as 'horn' than with one such as 'chair' (phonologically unrelated to target). The time course of phonological facilitation is exactly predicted by Roelofs's WEAVER++ model.

Upon retrieval of the code from the mental lexicon, the next operation is initiated, syllabification. Segments are incrementally concatenated to form syllables. Concatenating the segments h, o, r, s produces the phonological syllable /h o r s/. But if the target word would have been the plural noun (for instance when there were two horses on the picture), a disyllabic syllabification would have resulted: /h o r / – / s w z/. Hence, syllabification is context dependent. Whether the syllable /h o r s/ or /h o r / will be generated depends on the following context. Segmental concatenation in syllabification runs at a rate of about 25 milliseconds per segment. Incremental syllabification predicts a word length effect, confirmed by recent experiments: naming latencies are shorter for monosyllabic than for disyllabic words.

The final step in form encoding is phonetic encoding, the retrieval of articulatory scores for each of the incrementally generated syllables. The theory assumes the existence of a mental syllabary, a repository of syllabic gestures, motor programs for frequently used syllables. There is evidence for the involvement of premotor cortex/Broca's area in the storage of these overlearned syllabic gestures. The factual execution of these gestures by the laryngeal and supralaryngeal articulatory system generates the overtly spoken word. But that is beyond the present theory. ●


LEXICAL SELECTION
perspective taking
lemma selection
FORM ENCODING
retrieving phonological codes
syllabification
phonetic encoding
articulatory score