



Man and machine, deep in conversation: what was still an intellectual game in Stanley Kubrick's science fiction film "2001: a space odyssey" is now reality. Computers which understand human communication even better are due on the market in the next few years. Linguists have discovered that machine speech recognition operates according to the same patterns as in man.

# When **Computers** obey our every word

*Speech only works because of certain probabilities and rules understood equally by both speaker and listener. Only with the help of this underlying pattern can man combine individual sounds into words and thereby identify the meaning. This was the central thesis of the 40 or so scientists who attended the "Speech Recognition as Pattern Classification" workshop in Nijmegen in the Netherlands at the invitation of the MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS.*

**"C**omputer, tell me all there is to know about planet earth!" In the TV series "Starship Enterprise" Captain Kirk issued verbal commands to his instruments. Mouse and keyboard were superfluous. It is not just television producers and film makers like Stanley Kubrick, whose legendary creation "HAL" in "2001: a space odyssey" could lip-read human speech and, to the chagrin of the astronauts, soon developed a life of its own, who have fallen under the spell of intelligent computers which can understand human speech and have no problem conversing with people.

PHOTO: DEFEA-MOVIES

What was long considered a bold vision is gradually becoming reality. Not only researchers, but commercial companies are interested in automatic speech recognition and speech synthesis, or artificial speech generation. Much is expected of the new technology: only recently Microsoft's founder Bill Gates proclaimed speech recognition to be the "future of the computer". Yet, despite all the euphoria about opportunities for growth and gains in knowledge, scientists are forced to admit that machines are still a long way from human speech competence.

Until computers have "HAL's" intelligence and understanding, machines will need to be trained by humans. The astonishing thing is that they have to learn precisely those speech systems which infants acquire with apparent ease at the age of just a few months. The first things which small babies learn to understand are speech sounds which form simple words like "Mama" or "Dada". Children soon unconsciously recognise that the pronunciation of these elementary utterances is based on certain speech patterns. Even if it is not the child's mother, but a stranger, who pronounces a word, children can recognise the word "Mama" – regardless who it is spoken by. A pattern like this ("M-a-m-a") has to match all possible forms of pronunciation, regardless whether the speaker has a Brummie, Scouse or Geordie dialect – provided the dialects are not too extreme. Otherwise the acoustic signals cannot be translated into meaningful words and sentences.

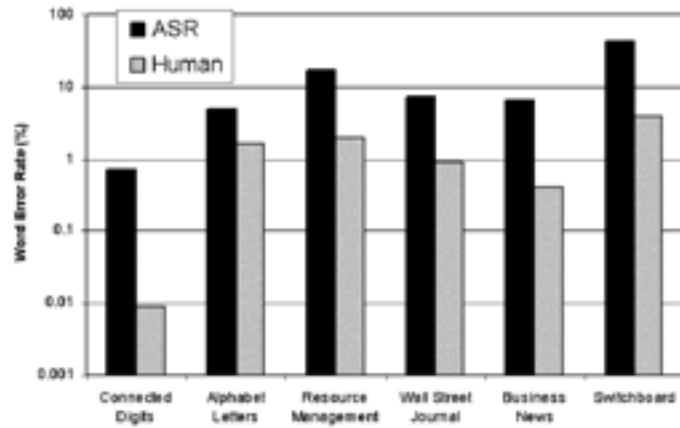
Even an apparently straightforward discovery such as this presents huge problems for scientists working on automatic speech recognition. In today's globalised world our language alters daily; at the same time, it is becoming increasingly important to communicate rapidly with one another and to overcome language barriers. Human and machine speech recognition plays a major role here if we want to know how speech recognition actually works.

Consequently, around 40 scientists from widely differing disciplines spent three days discussing, for the first time, the theme "Speech Recognition as Pattern Classification" in Nijmegen in the Netherlands at the invitation of the Max Planck Institute for Psycholinguistics. The question on everyone's lips before the conference began was: is there even any common ground between linguists, psychologists, engineers, technicians, and computer scientists who are all dealing with various aspects of speech recognition?

## ALL LEARN FROM ONE ANOTHER

The results of the conference in Nijmegen were encouraging. Scientists dealing exclusively with automatic speech recognition can learn a lot from their colleagues who specialise in human listeners. The reverse is also true. The psycholinguist Roel Smits, who is undertaking research at the Max Planck Institute in Nijmegen and helped organise the meeting, summarised the conference with the words: "There has been a great deal of specialisation in speech recognition research since the eighties. Now discussions between the disciplines have got off the ground again."

It became clear in Nijmegen that the scientists share a common basis. The researchers agree that speech only works if certain probabilities and rules exist, which are understood equally by both speaker and listener. Only with the help of these rules can the recipient of verbal utterances combine individual sounds into words, thereby identifying the meaning. For anyone acquiring language, the step from pattern to a category is the result of numerous instances drawn from past experience. This applies both to man and to machines. Small infants have to hear acoustic signals time and again until they are able to understand words and later whole sentences and, at some stage, independently form their own: they learn to assign individual patterns ("b", "a", "l") to a certain category ("ball"). This ability



Linguists from the Max Planck Institute for Psycholinguistics have compared the error rates in various types of text and discovered that human listeners are still far superior to machines, even with relatively simple tasks such as recognising series of numbers.

is based on probabilities, according to which children learn to form categories. “Man is a natural statistician”, says Roel Smits.

**COMPUTERS ARE FED HUNDREDS OF HOURS**

A speech recognition computer behaves in much the same way. In order to be able to recognise linguistic patterns and categories, a machine has to be “fed” a certain amount of data. This requires many hundreds of hours’ worth of recordings which store the linguistic utterances of different speakers. Ultimately, the machine should be capable of doing precisely what a person does each day: filtering out one particular speaker from the babble of voices in a department store. Or understanding the mumbled words of a conversation. And what about a whispered declaration of love? The computer must be able to recognise this, too, if it is to be a match for humans. In the race between man and machine, man still clearly has the advantage: “Listeners quickly adapt to a change in environment. Machines, however, are sensitive to a change in speaker or a different speaking rate. They are too easily distracted by background noise”, says the psycholinguist Smits.

Machines which understand the natural flow of words and linguistic peculiarities such as accents or dialects are already available, yet there is room for considerable improvement. There is definitely a market for such programmes: American busi-

nesses are already attempting to reduce costs and save on staff by using automatic speech recognition. Once such voice systems operate smoothly, expensive call centres will be a thing of the past and the telephonist, an early 20th century symbol of progress in communications, will finally be history. Instead speech recognisers would process customers’ detailed requests, take orders, answer questions, handle complaints – never once losing their patience.

The first systems, based on automatic speech recognition, already exist: anyone wanting to make an en-

quiry about a train connection in the Netherlands speaks to a machine which responds to the enquiry with a friendly voice. Only if the automatic enquiry has not succeeded by the second attempt, does a human assistant intervene. The dialogue between man and machine obviously only works between very fine linguistic lines.

**COMPETITION BETWEEN WORDS**

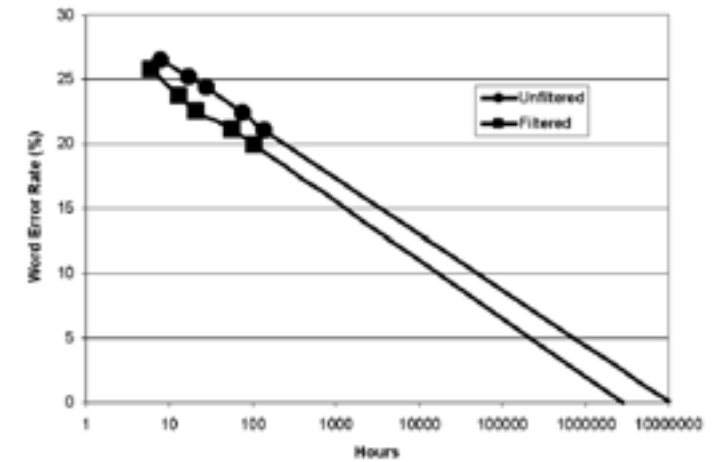
Yet basic research is still necessary, as the conference in Nijmegen made clear. For machines can learn a lot from the flexibility of human listeners. An example of this was provided by the linguist Gareth Gaskell of the University of York. Gaskell described how adults add new words to their vocabularies. According to his assumptions, speech recognition in man works in much the same way as in a digital lexicon which operates with a search engine. If a listener hears only the beginning of a word, all the words beginning with this sound are activated simultaneously – they compete with one another at this early stage. The syllable “cap” activates the words “cap”, “captain”,

“capital”, “capture”, “capsule”, etc. These words are only deactivated if the input no longer matches the stored form. If a “t” follows the syllable “cap”, then only “captain” and “capture” remain activated. This whole process takes just fractions of a second in the listener’s brain. The pool of possible words becomes progressively smaller until the right word is identified. In principle, speech recognisers operate in a similar fashion: in speech recognition they calculate the probability that a particular sequence of sounds will produce a word. As a result, a large number of lexical variants are rejected and ruled out in a continuous process.

Yet how are new words introduced into this lexical “competition”? What happens if a new word begins with the same sound as an existing one? Gareth Gaskell discovered that there is initially no competition, provided the listener does not hear the new word repeatedly day after day. The existing vocabulary only adapts gradually to new input – probably to protect previously learnt words from excessive interference. James McQueen and Anne Cutler of the Max Planck Institute for Psycholinguistics, however, found evidence that listeners keep flexibly adapting their criteria for identifying speech sounds, if circumstances so dictate.

**LISTENERS ARE TOLERANT AND FLEXIBLE**

A test series in the language laboratory brought definite proof: half the test subjects heard an ambiguous sound between an “s” and an “f” which occurred at the end of a string of sounds such as “kara-” which resembles the Dutch word “karaf” (carafe). They also heard distinct “s” sounds at the end of words such as “karkas” (carcass). The other half of the test group heard the opposite: a distinct “f” in words ending in “f” such as “karaf” – and the ambiguous sound at the end of the string such as “karka-”. The contrast between the ambiguous “s” or “f” sounds and the listeners’ knowledge of Dutch words



The perfect speech recognition computer does not exist – yet, with a lot of practice, the error rate will drop continuously. According to this calculation produced by Anne Cutler and Roger Moore, the programme has to be fed ten million hours’ worth of data to reach an error rate below one percent in speech recognition. The top line shows the experiment using unfiltered data from everyday communication. In an additional test, training material was selected and the error rate was consequently lower.

led the first group to hear “f” while the second heard an “s”.

The result shows that a radical change in pronunciation of a speech sound which distinguishes one word from another does not present the listener with a problem, provided he obtains sufficient information to place the sound within an existing word. Several of the Nijmegen papers made it clear that listeners tolerate varying pronunciations of existing words and adapt to new linguistic utterances. They have already heard so many different pronunciations that they can easily create a new association between a signal pattern and a stored category.

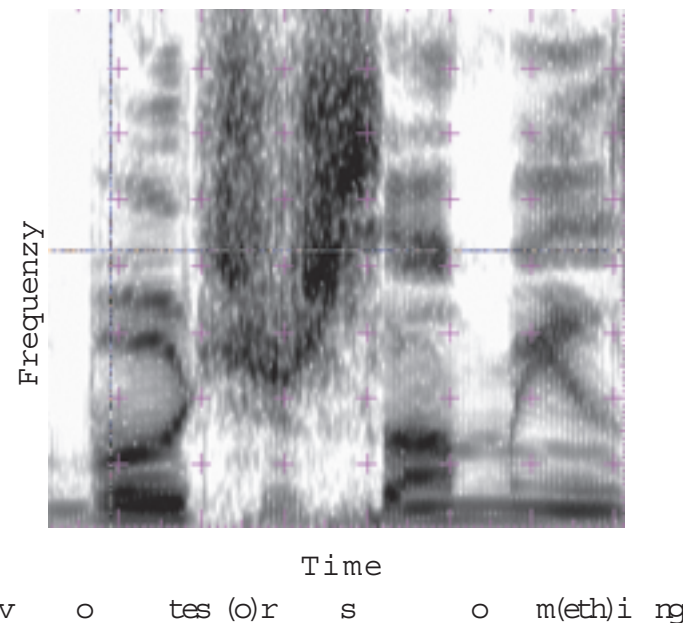
To find out more about the nature of speech recognition, scientists must examine the system in its early phase. Children learn at a very young age how to organise speech signals according to pattern and categories. Linguists can test the reactions of babies by observing the way test subjects suck on a dummy. To do this, the infant is exposed to various sounds in the language laboratory. If the baby understands the change in the sequence from “da da da” to “pa pa pa”, it automatically sucks faster. The child’s reactions can be measured electronically using complicated experimental apparatus. In a dif-

ferent experiment, the children sit in front of screens on which words like “boy” or “ball” appear. If the correct word is spoken by a speaker, the researchers can tell from the direction of the child’s glance whether they have understood the word or not.

**FAMILIARITY INFLUENCES LEARNING**

Children become receptive to different sounds between the age of six to twelve months. They can understand the sound of the words – provided the speaker is known to them. Problems arose in the experiment whenever the words were spoken by different people. If a word the child already knows is immediately repeated by a second speaker, then even seven-month-old infants are able to understand it. The effect of repeating it has worked. Experiments also reveal, however, that if there is an interval of one day between utterances of the same word, the child is no longer able to recognise the word because it does not know the second speaker. It is connected to a quite specific voice.

So people learn in early childhood to decode speech signals and to combine these signals into a meaningful whole. It takes a number of years to solve the puzzle. Speech recognisers



Speech recognisers record human utterances spectrographically. The three words “votes or something” are recognisable by their frequency, the dark areas indicate the vowels. However, the speaker has only pronounced the words cryptically, as shown by the spectrogram: “votes ‘r som’ing”. Slurring such as this presents a huge challenge to speech recognition programmes. The computer is also sensitive to irregular pauses in speech.

FIG.: MPI FOR PSYCHOLINGUISTICS



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.

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 adjust-

ed 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.”

PHOTO: CORBIS - STOCKMARKET

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 dis-

tracter 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 supra-laryngeal articulatory system generates the overtly spoken word. But that is beyond the present theory.

