HOW DOES THAT WORK?

SPEECH RECOGNITION

Speech recognition is a machine or program's ability to recognise spoken words and phrases and convert them into a machine-readable format. The software is now a common feature in several devices, including smartphones, computers and virtual assistants.

Speech recognition is an intricate area of computer science, using a mixture of complex linguistics, mathematics and computing. It has been revolutionised in the last decade or so by the application of artificial intelligence (AI) and is by far the largest current application of AI.

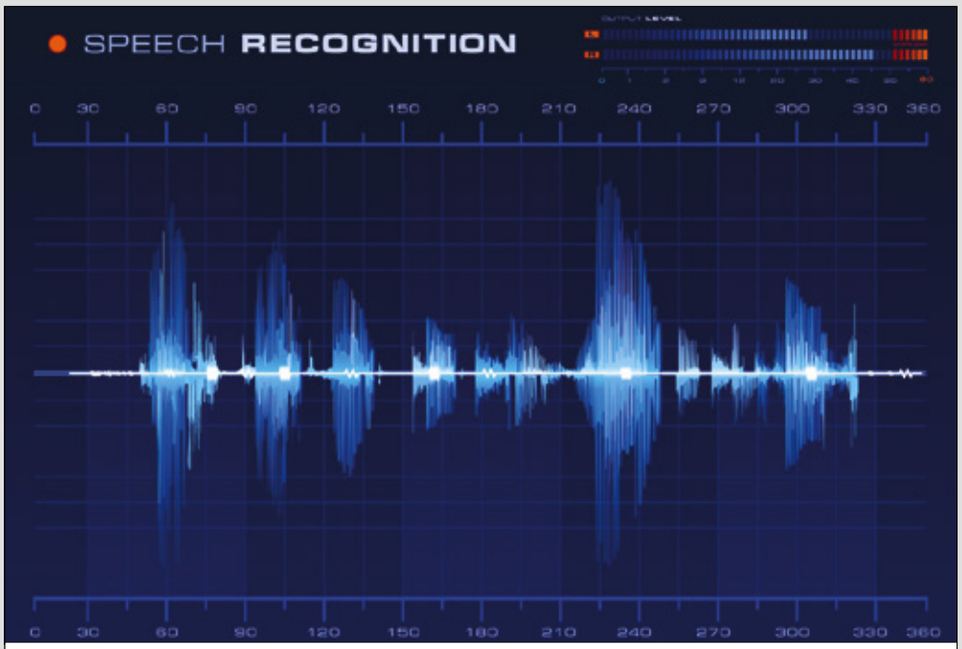
In simple speech recognition software, such as automated telephone systems used in call centres, the computer is trained to recognise a very small number of words, such as yes, no and numbers. It matches the sounds to preloaded patterns, and can recognise it through a range of accents.

Nowadays, computers carry out several steps to recognise human speech by digitising and processing sounds that can be matched to phonemes, the smallest unit of sound in speech (44 in English). These can be analysed to recognise them as meaningful language.

Speaking creates vibrations in the air that a microphone changes to a continuous electrical signal. An analogue-to-digital convertor converts the speech into a digital signal. It digitises the sound by taking measurements of the soundwave at frequent intervals and turning them into a digital format.

The computer processes this digitised signal to find the speech within all the captured sound; breaks it down into 'phones', small units of the actual sound, and processes these 'phones' to make them easier to compare to phonemes.

Any sound, and speech is no different, is made up of many frequencies just as a chord in music is made up of several different notes. The first two steps use signal processing techniques that identify the frequencies and their relative intensities at a point in time. Complex statistical models, and more recently AI, are used to identify the patterns within these that are speech and the 'phones' it is made up of.



Speech recognition software picks up the vibrations in the air created by speech and converts them into digital signals © Adobe Stock

The third step is to make those 'phones' consistent. When we speak, we speed up and slow down and the volume of our voice varies. To match 'phones' to standard phonemes the 'phones' are normalised – matched to a consistent rate and volume.

The program then needs to put each phoneme into the context of the other phonemes around them, allowing the computer to work out what it was likely that the user was saying. This is where AI comes in: training and statistical models help the speech recognition program recognise words that sound the same, such as 'see' and 'sea'. The context generally allows the program to work out which one is being used.

The AI task of recognising words correctly in the presence of background noise and different accents and individual speech patterns is considerable. Therefore, it is a task that cannot easily be carried out by laptops or smartphones. Popular speech recognition systems, such as those from Apple and Google, depend on passing the task of recognition to very powerful computers in the 'cloud'.

Research on analogue and digital electronics that might enable speech recognition in portable devices is ongoing but is still at an early stage.