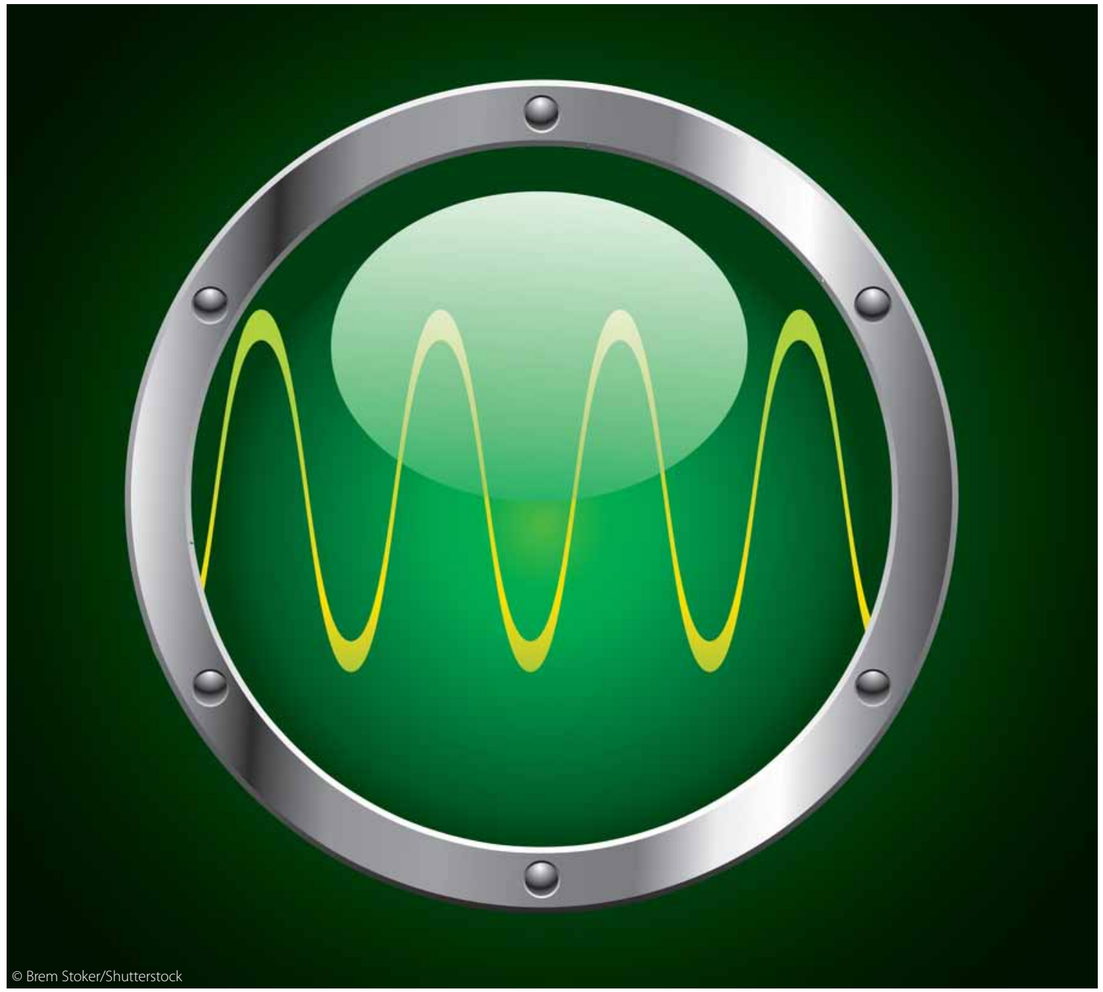
TALKING TO MACHINES



© Brem Stoker/Shutterstock

With the introduction of voice-driven personal assistants such as Apple's Siri and Google Now, speech recognition appears to have finally made it into the mainstream. Professor Steve Young FREng of the University of Cambridge's Information Engineering Division, reviews the progress over the last few years that has made these applications possible and considers the impact that future development will have on our ability to communicate with machines.

Speech recognition programs have become an increasingly large part of our daily lives. Paying for parking meters by phone and being guided to selected company departments by an automated voice is becoming more common. Voice-controlled interfaces can now be found in an increasing number of environments: mobile phones, televisions, and even cars.

There are various software products that allow users to dictate to their computer and have their words converted to text in a word processed or email document. There are some very successful programs that have been developed for specific business settings, such as medical or legal transcription. People with disabilities that prevent them from typing have also embraced speech recognition systems.

The core technology that makes this possible is automatic speech recognition (ASR). This is the process whereby the speech waveform captured by a microphone is automatically converted into a sequence of words. Given the sophistication of modern pattern recognition technology, this might seem to be a relatively simple task. However, in spite of the major progress that has been made over the last decade, there is still quite a way to go before speech recognition will be 100% reliable.

SPEECH RECOGNITION

Speech is made up of a sequence of words where each word consists of a sequence of basic sounds, which speech technology engineers refer to as 'phones'. In English, about 40 phones are required to construct every word in the language. With the phrase "I need a hotel", for example, the word "need" consists of three phones */n/*, */ee/*, and */d/*. Speech recognition is difficult because the choice and realisation of each phone is variable. The same words spoken by different speakers can vary dramatically, and even the same speaker will pronounce the same phone differently in differing contexts. For example, the */ee/* sound in "need" is acoustically different from the */ee/* sound in words such "seem", and "keel". Phones also change with speaking style (fast versus slow, casual versus dictation), mood and physical state (such as having a cold).

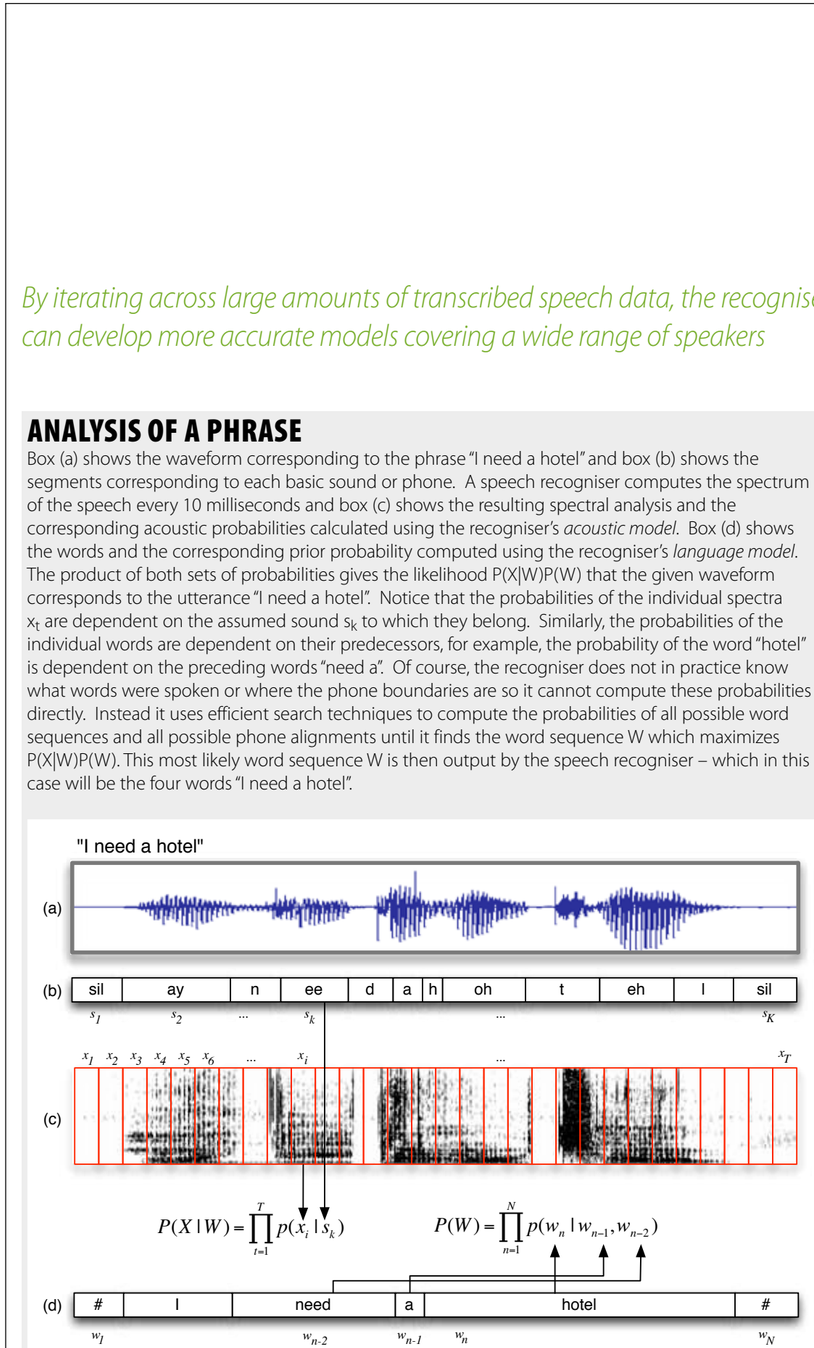
For the receiver, background noise and varying channel characteristics add further confusion. In addition, there are no acoustic cues which signal phone or word boundaries. So speech recognisers not only have to classify each individual sound, they also have to find the location of each sound in the waveform.

Human listeners effortlessly decode these potentially confusing sequences of sounds by exploiting their knowledge of vocabulary, syntax, semantics and common-sense reasoning. In contrast, automatic speech recognisers' knowledge is represented in the form of two probability distributions: an 'acoustic model' which provides the likelihood that an utterance corresponds to a given word sequence and a 'language model' which provides the prior probability of what is said. The acoustic model is composed of a set of distributions defining the probability of every possible sound/phone spoken in every possible context and the acoustic likelihood of a matching word sequence is formed from the product of the probabilities corresponding to each of the constituent phones. The language model is composed of a set of distributions defining the probability of every possible word given its immediate predecessors and the prior probability is formed from the product of the probabilities of each actual word in the given sequence. The problem of actually recognising speech is then reduced to the problem of finding the word sequence that

By iterating across large amounts of transcribed speech data, the recogniser can develop more accurate models covering a wide range of speakers

ANALYSIS OF A PHRASE

Box (a) shows the waveform corresponding to the phrase "I need a hotel" and box (b) shows the segments corresponding to each basic sound or phone. A speech recogniser computes the spectrum of the speech every 10 milliseconds and box (c) shows the resulting spectral analysis and the corresponding acoustic probabilities calculated using the recogniser's *acoustic model*. Box (d) shows the words and the corresponding prior probability computed using the recogniser's *language model*. The product of both sets of probabilities gives the likelihood P(X|W)P(W) that the given waveform corresponds to the utterance "I need a hotel". Notice that the probabilities of the individual spectra x_t are dependent on the assumed sound s_k to which they belong. Similarly, the probabilities of the individual words are dependent on their predecessors, for example, the probability of the word "hotel" is dependent on the preceding words "need a". Of course, the recogniser does not in practice know what words were spoken or where the phone boundaries are so it cannot compute these probabilities directly. Instead it uses efficient search techniques to compute the probabilities of all possible word sequences and all possible phone alignments until it finds the word sequence W which maximizes P(X|W)P(W). This most likely word sequence W is then output by the speech recogniser – which in this case will be the four words "I need a hotel".



X = utterance, W=word sequence, P(X|W) = acoustic likelihood, P(W) = prior

maximizes the probability of the product of the acoustic likelihood and the prior probability – see *Analysis of a phrase*.

MODEL DEVELOPMENT

These models are surprisingly effective. Their strength lies in the fact that they can both be *trained* automatically from data. Given a large database of utterances spoken by many speakers and the corresponding word level transcriptions, an automatic speech recogniser can easily find the location of the phone boundaries and then use the speech vectors aligned to each phone to update the parameters of the acoustic model. Thus, by iterating across large amounts of transcribed speech data, the recogniser can develop more accurate models covering a wide range of speakers. Similarly, given a large text archive, the probability of any word given its predecessors can be estimated by counting the number of times that the word occurs with those predecessors in the archive.

These basic elements of an automatic speech recogniser were established more than 30 years ago. However, achieving acceptable performance on natural speech has proved to be a significant engineering challenge. To cover the nuances of language, the acoustic model must consider each of the 40 or so phones in around 1,000 different contexts resulting in nearly 10 million parameters in total. These phone models must be robust in order to tolerate extraneous noise and adapt automatically to speaker-specific variations, requiring complex mathematical modelling.

Unrestricted vocabulary systems are typically built up using approximately 1,000 hours of speech, equating to 10,000 spectral training vectors per phone model. If each phone model is trained separately, then it is relatively simple to split the development over large arrays of computer servers to achieve acceptable throughput. However, the most recent systems are trained 'discriminatively' which involves training multiple models in parallel and forcing the system to choose one in preference to all of the others. This simultaneous training means that each phone model requires access in principle to all of the data at once. It is much harder to run systems in parallel and this has led to the recent trend to exploit banks of graphical processing units to achieve the necessary throughput.

Building a language model is equally challenging. A typical model will have around 100 million parameters and it will require at least 1,000 million words of representative text to train. Current performance has been achieved therefore by combining sophisticated machine learning techniques with large-scale software engineering. The accuracy of speech recognition is measured by the number of mistakes made for every 100 words analysed. Typical word error rates are now between 3 and 8% for clean, carefully spoken speech, rising to 20% or greater for conversational speech in noisy environments.

ROLE OF THE INTERNET

Until recently, speech recognition has been limited to relatively few languages (primarily English), and rather specific applications such as medical dictation where it is relatively straightforward to collect sufficient representative training data and provide the computing power needed to run the recogniser. Dictating directly onto a patient database,

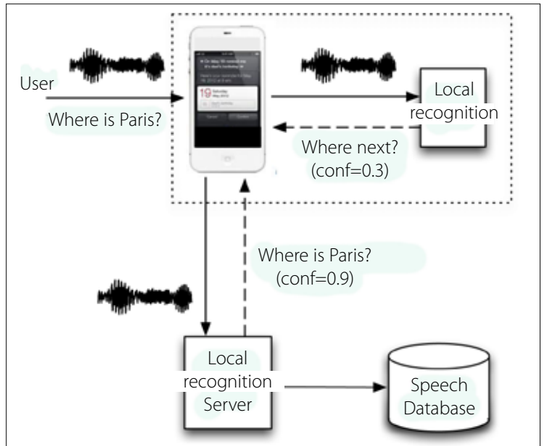


Figure 1 – Client-server recognition architecture. User speech is recognised locally and remotely. If the local recogniser is confident in the result, the remote transcription request is aborted. Otherwise, the phone waits for the server to reply. When 'conf' (confidence) =1, the recogniser is certain. Confidence scores are used to avoid the system being confused by transcriptions which have many errors in them. In either event, the speech can be stored in a database and used to iteratively update the remote server's recognition models

Voice recognition software in apps and mobile phones allows users to search and interrogate the web © Peter Da Silva

rather than via recording equipment and a secretary, saves about a day in updating records. Now, the rapid growth and speed of the internet is changing the way speech-based systems are deployed and expanding their potential to embrace new languages and applications.

Consider the typical configuration of a personal assistant application running on a smartphone. If the user asks the question "Where is Paris?", the waveform is passed to a small local recogniser running on the phone which is configured with a modest vocabulary to answer common questions relating to the built-in apps such as phone, contacts and calendar. At the same time, the waveform is sent to a remote server running a much more powerful large vocabulary recogniser. If the local recogniser is confident in its response, then the transcription request to the remote server is aborted since it is no longer needed. If trust in the local recognition is low, then the phone waits to receive the reply from the server – see *Figure 1*.

This organisation has a number of advantages. It allows fast response to common user requests while at the same time it avoids the limitations and frustrations of poor and limited speech recognition capability. The ability to transmit audio around the planet with minimal latency is a relatively new phenomenon and it is this that makes the client-server architecture shown in Figure 1 viable and popular.

Another advantage of this architecture is that every speech waveform which goes to the remote server can be analysed and stored in a database. The data can be then used to improve the performance of the system. Thus, a supplier can offer a service in a language or dialect for which there is little training data, and then rapidly collect data from users and repeatedly retrain the system. Modern active learning techniques allow training to use a mixture of hand-transcribed and automatically transcribed data. The former used to be very expensive to produce, but here again the ubiquity of the internet has provided a solution. With a well-designed web interface, crowdsourcing sites such as Amazon Mechanical Turk can now provide a way to get simple repetitive tasks such as high-volume data annotation performed very quickly at low cost. The net result is that users see a rapid improvement in performance while the costs to the supplier of data collection and system improvement are much reduced.

UNDERSTANDING SPEECH

The ability to translate speech into digital text enables users to dictate messages and search the web and these are now standard features of both iPhone and Android smartphones. However, more sophisticated applications require the ability to actually understand the meaning of the words. In order to respond to a spoken command such as "Arrange a meeting with Mike Monday at four" it is necessary to identify "Arrange a meeting" as being the action, "with Mike" as being an attendee and "Monday at four" as being the time. Since the recogniser cannot determine capitalisation from the acoustic signal alone, the attendee could also be someone called "Mike Monday" and the time could be "at four" implying at 4pm today. Understanding speech therefore requires decoding the phrasal structure, mapping the phrases to actions and entities, and resolving ambiguities.

This is done by using statistical techniques which automatically apply mappings from words to application level semantics. These systems have the added benefit of resolving ambiguous interpretations by simply selecting the most frequently occurring option in the training data. Thus, in the example above, a statistical understanding component would probably identify Monday as part of a time simply because the phrase "Monday at …" will be very common in the data – see *Statistical speech understanding*.

As with speech recognition, large amounts of data are required for training statistical classifiers. However, if the application runs on a client–server architecture, user utterances can be collected and then manually annotated by crowd-sourced transcribers. For understanding more general queries such as "Find me a movie by the director of Titanic", web queries provide an alternative source of training data because users usually respond to the results of a typed search query by clicking on the most relevant result. Given sufficient numbers of these query-click pairs and access to an ontology (a tree structure defining the semantic relationship between words) such as DBpedia.org, it is possible to automatically learn mappings from phrases to actions and entities.

SPOKEN DIALOGUE SYSTEMS

Whether rule-based or data driven, the combination of speech recognition and speech understanding provides all that is required to implement the current generation of smartphone-based personal assistant. The user speaks to the device and the system attempts to understand what is said. If it can interpret the utterance as a command, it executes it; otherwise it passes the word string to a search or query engine which operates exactly the same as if the command had been typed.

The current generation of personal assistants needs to be improved. If words are misrecognised or the user's intention cannot be understood, the interaction fails. These personal assistants are really just input devices; they do not collaborate with the user to help achieve their goal. As illustrated by the following example, collaborative problem-solving requires a dialogue both to understand the user's requirements and to seek an acceptable solution.

User: Find me somewhere to get some lunch.

System: Any preferences?

User: Thai would be good, otherwise Chinese.

System: The nearest Thai restaurant is in Soho, but you have a meeting at 2pm.

User: Hmm, thanks for the reminder. Is there anything nearer?

System: There is a reasonable Chinese restaurant about 5 minutes away.

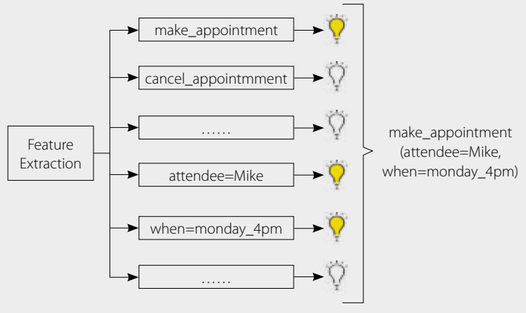
User: Ok, book a table and give me directions.

Spoken dialogue systems extend basic speech understanding by including a mechanism for tracking the user's beliefs and intentions with a decision component to determine how to respond to the user in order to help achieve the required goal. Current spoken dialogue systems are rule-based and are limited to very simple applications. They are very fragile in the face of recognition errors and have to be laboriously programmed for each new situation. As with understanding, the solution is to move away from rules, to learning from data. However, this is not a straightforward

STATISTICAL SPEECH UNDERSTANDING

A classifier, typically a support vector machine, is created for every possible semantic element and trained to recognise that element whenever it appears in a spoken phrase. The speech is converted to a set of features such as a vector containing the counts of all 1,2, and 3 word sequences (called N-grams) in the utterance, and these features are then input to a bank of classifiers. Each classifier is trained to recognise a single unique action or entity. When an utterance is input to this system, the set of positive classifier outputs are combined to construct the required semantics.

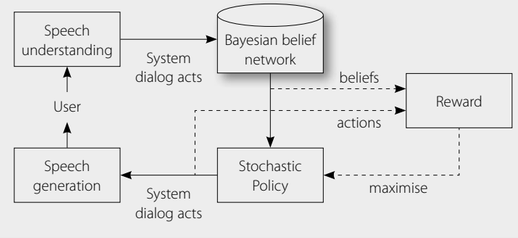
This type of brute force approach is surprisingly effective. Not only is it robust to speech recognition errors, ungrammatical input and ambiguity, it can also learn to correct commonly occurring errors. If the training data contains examples of the error along with the corrected output, the relevant classifiers will learn to output the correct semantics even when the error occurs.



The ability to translate speech into digital text enables users to dictate messages and search the web and these are now standard features of both iPhone and Android smartphones.

A SPOKEN DIALOGUE SYSTEM USING REINFORCEMENT LEARNING

The user's beliefs and intentions are modeled probabilistically using a Bayesian belief network. For example, in a restaurant information system, the network would comprise of a random variable for each possible attribute of a restaurant such as food type, price range and location. As the user discusses the type of restaurant they are looking for, the probabilities of each of the possible attributes are updated using Bayesian inference. The system's responses are generated according to a policy which is a stochastic (statistical) mapping from beliefs to actions. The system receives positive feedback from the user and adjusts the mapping to maximise the reward. Over time the system learns to react optimally even when the speech recognition accuracy is poor.



pattern recognition problem; rather, it is a problem of planning under uncertainty and it requires a branch of mathematics called 'reinforcement learning' to solve it – see *A spoken dialogue system using reinforcement learning*.

A telephone-based system has recently been built using this approach at Cambridge University to help users find restaurants in the city. Starting from scratch, the system learned how to respond at every turn by interacting with users and asking them at the end of each call whether or not they were successful. By giving the system positive feedback for success and negative feedback for failure, learning progressed rapidly and the system achieved a 97% success rate after 1,000 dialogues, which is comparable to a human operator.

INTEGRATING THE INTERNET

Recent developments in machine learning and the emergence of smart devices connected to a global high-speed network open the door to much richer and more accessible interaction with machines and information. The key will be to build systems that learn and adapt from experience. Once the need for human experts to hand-craft applications for each new domain and language is removed, the roll-out of speech-based interfaces will accelerate. Core speech recognition performance will continue to improve incrementally as statistical models of acoustics and language improve and training data sets grow ever larger. However, the key engineering challenge is to find effective ways to draw together the vast quantities of data embedded in the internet so that machines can associate meanings with the words we speak and learn to use the power of natural conversation to explore and exploit information.

Future generations of personal assistant will then do much more than provide an alternative to typing, tapping and swiping. They will become more and more intelligent, able to hold conversations, recognise tones of voice, learn the user's likes and dislikes, provide information and undertake simple tasks such as organising meetings, ordering goods and chasing customers for payment.

Ultimately the social impact may be profound. For example, specialist automated health-care assistants could give personalised guidance to the elderly and infirm, greatly reducing the burden on national health systems. Advice will be on hand to anyone on any topic; all they will have to do is ask.

Listen to:
www.bbc.co.uk/programmes/b01phlgn

BIOGRAPHY

Professor Steve Young FREng has worked in the area of speech recognition and spoken dialogue systems for more than 30 years. He invented a software toolkit for building speech recognition systems called HTK which is widely used in both academia and industry. He received a Technical Achievement Award in 2004 from the IEEE Signal Processing Society and the 2010 Medal for Scientific Achievement from the International Speech Communication Association.

Toyota includes voice recognition technology in its most recent vehicles, accessible via the steering wheel. The system allows natural speech rather than needing to memorise and recite specific preset commands. It is linked to various multimedia functions including satnav and phone © Toyota