Table of Contents

TABLE OF CONTENTS ........................................................................................................... 2
INTRODUCTION ...................................................................................................................... 3
HOW IT WORKS ...................................................................................................................... 4
COMMAND AND CONTROL VERSUS LARGE VOCABULARY SYSTEMS ......................... 5
GETTING GOOD RESULTS ..................................................................................................... 7
REAL-WORLD APPLICATIONS .............................................................................................. 8
ABOUT THE AUTHOR ........................................................................................................... 10
Introduction

Speech. How simple it seems to most of us. Spoken communication is such a fundamental part of human life that we rarely stop and think about how amazing an activity it really is. The vocal system is a very complex part of our anatomy and producing spoken language relies on highly sophisticated, and still not fully understood, physical and mental processes. Understanding spoken language is possibly even more elaborate. Yet, as we go about our daily lives, most of us remain blissfully unaware of these complexities. It comes, indeed, naturally to almost all of us.

Because spoken communication comes so easy to us and is such an elemental part of life, it is perhaps understandable that from very early on in the history of computing, we have sought to make our computers recognise speech. Indeed, speech recognition technology has been around for decades and today speech recognition engines are available for PC users as off-the-shelf software.

Yet, even the most recent speech recognition technology remains constrained by quite sharply defined limits and is by no means able to truly understand speech in the way humans do.

In this paper, I will provide an overview of the possibilities and limitations of speech recognition technology and its application potential.
How it works

Hidden Markov models form the basis of modern speech recognition engines, in fact have done so for decades (in fact the Hidden Markov approach dates back to the late 1960s). The reason for the longevity of this approach is that so far no really better methodology has been found. There is plenty of literature on Markov modelling and a detailed explanation falls outside the scope of this article (search for classic papers by Baum on the topic), but I will briefly highlight the core properties that make Hidden Markov models so persistent in the field of speech recognition.

Speech audio is a signal type with some particular properties: continuous, non-stationary and often not very pure (i.e. affected by noise as defined in information theory). Speech is essentially stochastic and therefore needs to be analysed using statistical probability methods. Hidden Markov models provide such statistical description of the speech signal. This analysis renders speech into a series of phonemes, which are combined using n-gram models. N-grams are essentially probabilistic models defining the most likely sequences of phonemes and words.

It should be clear from this description that, as already stated, a speech recognition engine does not really understand speech. The informational content, the meaning, of the speech is unknown to the computer. The probability models cannot resolve questions that rely on comprehension, or, even worse, require contextual knowledge to allow correct interpretation. For example, making the distinction between "I scream" and "ice cream" will require true understanding and appreciation of context, something computers can simply not do as they do not possess true intelligence in the human sense.

As such, speech recognition will remain severely constrained and no match for the human brain in the ability to understand spoken language. We have seen improved recognition rates over the last decades, not in the least because modern computers have more processing power and memory at their disposal (a result of Moore's Law), thus allowing a more in-depth, more finely granulated analysis and more cleverly designed text building algorithms. Extra capacity and performance means more opportunity for secondary inspection of the (partially) recognised speech. However, fundamentally speech recognition remains constrained by the unmovable barriers that lack of true understanding entail. The best hope for a real breakthrough is thus closely related to developments in Artificial Intelligence, and that is another field where science and technology have made very little true progress. We are still a long way from building computer systems that really understand information in the proper sense. In the mean time, it is vital that users of speech recognition systems, and designers and engineers implementing products and services based on speech recognition, are fully aware of the constraints and limitations of the technology.
Command and Control versus Large Vocabulary Systems

Since speech recognition basically uses statistical analysis to match spoken output to phonemes and then words and phrases, it will be obvious that recognising only a small set of fairly distinctive (in terms of statistical properties) words is simpler and can be more accurate than identifying thousands or tens of thousands of words in unconstrained speech. Voice control systems not only need to recognise only a relatively small set of words, the actual combinations of words in command sentences further limits the domain for any given phrase. Because of this, command and control style recognition can work even in less than optimal acoustic circumstances and without user specific training of the system. It is even possible to create voice and control systems for users with severe dysarthria (although user training will often be required in that case).

A good illustration of a command and control application is the voice control system I have in my car. It's called Linguatronic, a factory fitted system operated by pulling a lever and speaking commands such as "listen to phonebook", "dial number", "next track", etc. It allows me to operate the multimedia and sat-nav functions without taking my eyes of the road and I find it very useful indeed. Because the system only needs to recognise a few hundred words and only in a fairly limited number of combinations, it can work well even in a car driving at high speed and without speaker training (I speak with quite an accent in English as I am not a native speaker, but that does not bother the Linguatronic system).

By contrast, recognising large vocabularies across almost any topic is a much harder job. The differences between utterances can be very subtle and words
can appear in almost any context (especially if taking into account the fact that people seldom talk in a way that rigorously adheres to grammar and other formal language rules). Moreover, living language usage evolves all the time. But the true weakness of speech recognition systems is that they do not understand language in the way humans do. It means that for large vocabulary applications, the output inevitably will be less than fully accurate.
Getting good results

To achieve high recognition accuracy, especially for large vocabulary systems, a number of conditions must be fulfilled. These requirements relate to those factors that will influence the statistical analysis most, of which the nature and quality of the input signal is key. It is essential to use a clean, wideband signal. Consequently, using a good quality microphone that is positioned well so that no breathing and other noises interfere with the speech, is very important. Especially on laptops (where a lot of electronic components are situated very close to one another), the analogue audio circuitry can sometimes be prone to interference degrading the signal. Using a USB microphone (particularly one designed for use with speech recognition) is often better than using the analogue mic-in line on a soundcard (notably low-end ones) as it produces a cleaner signal.

A wideband signal is also very important, as it means that more features of the original signal are retained, allowing for better accuracy and higher precision during the analysis. That is why recognising speech from narrow band audio sources (such as a traditional telephone signal) produces far worse recognition accuracy compared to direct, wideband, high quality microphone input (and if such a signal is used, it is important to use the statistical models, corpora, that are derived from a representative narrowband signal instead of using wideband acoustic corpora).

Even if the input is of high quality, too much background noise (music, other people chatting, traffic noises, etc) will also heavily impact on recognition accuracy. So, the acoustic environment should be as free of background noise as possible. Many systems can be trained on specific user voices, and if the system is to be used in an environment that is by nature noisy, system training can also help in dealing with background noise, provided it is of a fairly stable nature, i.e. the background noise during training is representative of the usage conditions).

If the application is a simple command and control interface that needs to recognise a few dozen fairly distinct utterances (as in the case of the Linguatronic system I mentioned earlier), even a limited bandwidth input signal of only moderate quality might still be sufficient, but for larger vocabulary systems the audio input must be as clean and wideband as possible.

As already stated, speaker training of the speech recognition engine can also significantly enhance accuracy. This is because training produces a statistical model of the user’s voice and background audio, which can be taken into account when analysing the speech. Other factors that will have notable impact on accuracy are pace, intonation, articulation (accents) and making sure that phrases have a correct structure.
Real-world applications

Despite its limitations, present speech recognition technology can be a very useful tool for a variety of applications, as long as designers and users fully understand the boundaries and weaknesses of such systems. It is regrettable that the desire to hype up a new product or generation of speech recognition engine sometimes leads to blatantly misleading statements or misrepresentation of the realities of speech recognition and its role in real-world delivery. Notable examples in this category are the Lernout & Hauspie debacle, or more recently the Spinvox saga.

Speech recognition is already used for live subtitling on television, as dictation tools in the medical and legal profession, and for off-line speech-to-text conversion or notetaking systems. For all these applications, human editing of the output is needed to achieve really good levels of accuracy. In addition, and as already mentioned, there are an increasing number of small vocabulary or specialised command and control applications, from sat-nav systems and voice command in smartphones, to home automation.

What is not feasible with the current state of science and technology, is to produce a system that converts free, natural speech into text in a fully reliable manner or with at least human-level accuracy. In my previous role as a director of technology at an organisation for people with hearing loss, I often got the question why we didn't produce an app for a mobile handset that people with hearing loss could use to convert speech on the fly into text while going about their daily business. As should be clear from this article, that is well beyond the capabilities of current science and technology. Apart from the fundamental problem, covered earlier, that such a device would not truly "understand" the speech, the real-world situations in which it would need to function make it unfortunately quite unfeasible: the places where it would be most needed (ticket counters in a post office or the underground, meetings, receptions and other noisy environments, out and about in the city) are acoustically totally unsuitable environments for this application.

There are obviously other problems with the concept of large vocabulary speech-recognition on the move (and which is not to be confused with the small vocabulary command systems on mobiles already referred to), such as the limited processing power and memory of mobile devices compared to PCs. However, those limitations should ultimately disappear under the influence of Moore's Law. Also, with mobile broadband now already well advanced, it would be feasible to put the recognition function in the network and just use the mobile device to capture the audio, stream it to a network based recogniser and get the text back. However, the fundamental problem of non-understanding remains, as do the problems of background noise. Microphones in mobile handsets are also
often producing an inferior signal compared to a high quality microphone for
desktop usage.

**In conclusion:** speech recognition offers real potential, but comes also with
fairly rigid and significant limitations. Unless we make real progress in the field
of Artificial Intelligence, these limitations will broadly remain. Designers and
engineers building products and services based on speech recognition must
take the limitations into account and user expectations must be managed
appropriately.
About the Author

Guido Gybels is a veteran Information and Communication Technology (ICT) expert and senior manager, with a proven track record of delivering award-winning innovation, research and development activities, software and hardware engineering projects, standardisation and policy and regulatory strategy in ICT. A former Director of New Technologies and Director of Technology, he is an accomplished senior manager with in-depth experience of planning, reporting, line management and controlling large budgets.

Over the last two decades of his professional career, Guido Gybels has been at the forefront of digital technologies for desktop and mobile alike, with a special interest in Internet applications, usability and user-focused design. His ongoing role as a board member of the Centre for Usable Home Technologies (CUHTec) at the University of York (UK) is testament to his strong commitment to harness modern technology and new media to deliver innovative new solutions in the real world.

Guido combines strong technical knowledge and experience with well-developed leadership skills, strategic thinking, an evidence-based approach to management and an uncompromising commitment to excellence.

In addition to being a very accomplished expert and leader in ICT, Guido Gybels has also demonstrated great political and public relations skills. He has extensive experience as a media spokesperson, successfully delivering high profile media coverage in the broadsheets, on radio and television. He acts as an expert advisor to both the UK government and the European Commission, in which role he interacts directly with senior civil servants, MPs, Ministers and European Commissioners.

While being atypical for the profession, his educational background indicates an agile and curious mind, having studied and published in such diverse fields as history, linguistics, geography, didactics, docimology and psychology.