

# Successful speech and language processing - the challenges

A White Paper by Andy Walker, Head of Product Marketing, Vocalis

Until quite recently, claims for speech recognition products - most notably those used for dictation systems - could only be called extravagant. Such products simply did not live up to the claims made for them - that they could provide an interface to a PC which was not only more natural than the keyboard they were intended to replace, but also more efficient.

However, important advances have been made and there is little doubt that, in growing numbers of application areas, speech recognition is showing real signs of maturing. Take telecommunications, for example. Here, interactive voice response (IVR) systems, based on powerful advanced speech recognition (ASR) engines, can deliver excellent performance and provide network operators with the ability to develop service differentiators in an increasingly deregulated - and hence competitive - market. In fact, the technology is proving so successful within the telecommunications industry that network operators are constantly looking to enhance their IVR services through, for instance, the use of larger vocabularies and customer controlled updates to systems.

Despite the notable success of IVR within the telecommunications sector, there is still some way to go before the technology can be realistically said to have come of age. And even as speech recognition technology matures, there are many other issues relating to language processing in general that will continue to confront the R&D departments of IVR developers.

There are several contributory factors to the overall performance of a recognition system. Clearly, the fundamentals of the information technology industry – hardware processing power and advanced software functionality – are central. However, other slightly less obvious elements also play a crucial role, such as Human Factors, which describes the design and usability of systems. Applications built using state of the art hardware systems and software techniques will still perform badly if they are not engineered properly in terms of Human Factors.

## **Hardware**

Speech recognition is a performance-hungry process – it needs a lot of raw processing power. Hardware developments have played an important part in the significant progress that speech recognition has made in the last few years. But cost and space constraints mean it is vital to continue to optimise the performance of hardware, especially in the area of telecommunications. That is because network operators will be looking to continuously increase their call handling capabilities, and larger

numbers of speech recognition engines will be needed to handle the growing number of lines.

Within the semiconductor industry there is an established roadmap according to which microchips will continue to develop greater processing power for at least the next decade and probably beyond. Moore's Law – first enunciated by Gordon Moore of Intel – states that the number of transistors that can be put on a chip doubles every 18 months. Today's most advanced chips contain several million. The speed at which whole systems are developing, in terms of their memory capacity and processing speeds, are if anything out-performing even this rate of progress.

A projection by the general manager of Intel's Microprocessor Products Group, Albert Yu, says the typical processor in 2011 will probably have a clock speed of 10GHz, contain 1bn transistors, and process 100bn instructions per second. The implication is clear: if speech recognition is in any way constrained today by hardware limitations, those constraints will be removed rapidly.

Today, various architectures can support ASR. Digital Signal Processors (DSPs), for example, are a form of specialised microprocessor. However there is also advancement in the various architectures which leads to better memory interfaces and/or lower overall cost.

These new processors can also be mainstream, industry standard devices, thus giving a further benefit. This is that it gives customers the assurance that the supplier of their IVR system has support from multiple vendors, as well as providing a wide choice of devices and development tools, plus a large body of developers to support future enhancements.

## **Software**

The sort of clear-cut roadmap discussed above in relation to hardware does not exist for software. Advances here are much less easy to predict. However, for speech recognition, it is probably fair to say that it is in the software realm that the hardest challenges lie, and from which the greatest advances are likely to come.

For example, one challenge in building IVR systems is how to tell when someone has finished talking. The answer sounds obvious – wait for silence. In fact, it is one of the more perplexing problems in IVR because while our brains can easily discriminate between signal and noise, it is not easy to teach a computer to do so.

Consider a busy open plan office, for example, where background noise levels are likely to be high, and often made up of other people's speech. This scenario, an easy one for the human to deal with, is a very significant problem for a computer. The computer needs to detect the beginning and end of a caller's speech. Vocalis is promising a major step forward with new techniques in this area to help this process be quicker and smoother.

Further developments are greatly improving processing performance by handling intensive applications more efficiently. For example, one of the most popular uses for an IVR system such as Vocalis' *SPEECHtel* is to automate either wholly or to some extent a network operator's directory enquiries service. This inevitably involves the handling of much larger datasets, perhaps 1000 words, compared with applications that require the recognition of connected digits, where typically only the 11 words denoting zero and the numbers digits 1-9 must be recognised. Despite the hundred-fold difference between the size of the vocabularies, such applications require only three times more processing capacity for the larger datasets than for connected digit recognition – a highly significant reduction in the processor bandwidth required.

These techniques can be supplemented by still higher level processing. For example, by building a model of the structure of the expected dialogue, improved recognition accuracy and speed can be achieved. This exploits the fact that the probability of one word following a previous one varies significantly, depending on their functional/grammatical role in the dialogue. If the first word has been recognised and analysed grammatically, it becomes possible to reduce the possibilities of what the next word could be.

There is already a great deal of research work being conducted in this field of computational linguistics. However, much of the work has been done on text processing for use in applications like web search engines. Only relatively recently has human-computer dialogue been identified as an important area for research. In general, creating computer systems capable of more sophisticated language processing is seen by many as a potentially rich research area for the future, and may eventually lead to machines capable of powerful semantic analysis, leading to true natural language understanding.

Developments in speech and language processing are taking place on many fronts at once, as illustrated by the following key R&D topics. Although progress is being made, each of these areas will pose tough challenges for researchers well into the 21<sup>st</sup> century.

### **Speaker verification**

The ability to verify a caller's identity using ASR can help to combat fraud in card based applications in areas such as banking and telecommunications. Although there is still some way to go before this technique is perfected, real progress has been made very recently. Results of a research EU project, in which Vocalis' technology was central, have been encouraging, and the research is now continuing with a second project.

The first project, CAVE (Caller Verification in Banking and Telecommunications), culminated in tests of two Speaker Verification (SV) systems, a telephone banking system in Zurich, Switzerland, and a telecommunications calling card application in the Hague, Netherlands. Vocalis supplied the SpeechServer that hosted the SR and SV resources used in CAVE, and worked closely with several other partners,

including the Netherlands' PTT Telecom, Union Bank of Switzerland, Telia Research and France Telecom.

The final CAVE report says: "SV technology is now undoubtedly ready for deployment in well designed, low risk applications. In medium to high-risk applications, SV can be used in combination with other security measures. Embedding SV in existing human-machine interfaces might really lower thresholds to the effective use of tele-information and transaction services."

Ease of use and robustness, together with a single 'enrolment' procedure, were seen as critical requirements for SV-based access to telecommunications services. Enrolment is the initial process in which users input their voice so that the system can recognise on future occasions. For higher security applications such as electronic commerce, the report acknowledges that more powerful SV strategies are needed, but is optimistic this will happen.

Due to the success of CAVE, the work is continuing with a project called PICASSO (Pioneering Caller Authentication for Secure Service Operation), in which one of the central research targets will be adaptive enrolment, as Yong Gu, a Vocalis scientist, explains.

"The performance of SV systems has been shown to depend significantly on the amount of data used for enrolment – the more data the better. But lengthy enrolment procedures are not convenient for users. Thus, if a system adds to its enrolment data as it is used, performance can be continuously enhanced without that drawback. Also, people's voices change over time, so developing a system that can adapt to likely changes is important. The challenge is to develop a system which learns how the voice is changing over time."

### **Natural Language Parsing**

The ability to parse sentences from ordinary spoken language is a step on the way to solving one of the most challenging of all problems that will face the technologists of the 21<sup>st</sup> century: creating dialogue systems that can perform true, comprehensive natural language understanding. When this is achieved, the combination of powerful ASR and language understanding will result in machines having an entirely different level of capability to those of today.

Parsing is a process that can be carried out at several different levels, such as syntactic, or semantic. A syntactic parser operates essentially at the grammatical level and can analyse the phrase uttered to determine whether the string of words is a question, a statement or another kind of speech act. This is critical information in applications such as IVR systems, as it provides a clue as to whether the user is likely to have finished speaking an utterance and a response is needed. It also helps to constrain the semantic interpretation of the utterance.

A syntactic parser can be integrated with a semantic parser. Semantic analysis makes use of syntactic and domain-specific information to work out the meaning of an

utterance, at least in some limited sense. The output of a semantic parser is information regarding the meaning of the utterance that was (probably) spoken by the user of the system. One benefit of this higher level processing is that it can help to enable the system to decide when the utterance is 'complete' and therefore that a response is required.

Another reason for using semantic parsers is that they can make it possible to use dialogue history to deal with such language phenomena as anaphora. Anaphora describes the use of a word such as a pronoun that has the same reference as a word previously used in the same discourse. For example, in the sentence "John wrote the essay in the library but Peter did it at home", both 'did' and 'it' are examples of anaphora. To process them correctly requires knowing what went before.

Grammar descriptions are used not only to provide syntactic and semantic parsers with rules for the analysis of the strings of words passed on by the speech recogniser. They are also used to produce constraining grammars for the speech recognisers themselves. Use of such grammars not only improves recognition rates, it ensures that only syntactically correct strings of words reach the natural language parser.

Other forms of analysis play a role. For example, interpretation of prosody – the patterns of stress and intonation in language – can help to ascertain where a spoken sentence ends and what type of sentence it might have been. Simple prosody, such as the raising and lowering of tone, is normally used at the end of sentences and to mark questions or statements. Coupled with syntactic and semantic data, prosodic information can give strong clues as to the nature of an utterance.

## **Human factors and dialogue design**

Successful IVR products do not rely solely on speed or accuracy of response. Equally important is dialogue design, which should draw on wide-ranging research into language use and other 'human factors' expertise. This can make the difference between a system that delights callers, and a virtually unusable implementation, which may even create a negative effect on users and their perception of the entire organisation providing the so-called service.

For example, any IVR system that does not provide efficient fallback to a human operator runs the risk of leaving callers stuck in a 'voice jail' situation, from which the only escape is to hang up and try again – as the frustration mounts.

Human factors for IVR consists of several important areas including:

- Error correction and detection. IVR systems must be able to confirm that certain information has been correctly understood. Detecting errors can take the form of an explicit request, such as 'Did you say twenty-one?', or a more subtle repetition, such as simply 'Twenty-one?' (a case where the question contains an answer, known as an ellipse). Error correction can take various forms, ranging from the repetition of the statement in question through to moving to a different kind of dialogue structure.

- Dialogue initiative. There are various ways in which dialogue can be controlled: system directed, when the system asks questions that determine the dialogue flow; user directed, more likely to support more experienced users; or a mixture of the two.
- Making users aware of the IVR system's functionality. This is more difficult for speech based systems than for visual ones because vision is inherently permanent and parallel while speech is transient and serial.
- Output design – includes issues such as wording styles, speed, pitch, tone and degree of naturalness
- Cues for turn-taking
- Designing for novice and expert users – for example, allowing talkover and 'barge-in'

## **Research projects**

Many researchers around the world are tackling these issues. For example, an EC project called DISC (spoken language DIAlogue Systems and Components), in which Vocalis is participating, is aiming to establish a model of best practice procedures and methods for spoken language dialogue systems, their design and evaluation. This covers ASR, human factors and systems integration in order to establish dialogue engineering standards.

The goal of DISC is to develop a detailed and integrated set of development and evaluation methods and procedures for dialogue engineering best practice, as well as a range of support concepts and software tools. Ultimately, the methodology produced by DISC should help to establish dialogue engineering as a sub-discipline of software engineering.

The underlying aims of DISC are to minimise risk and improve procedures, methods, concepts and software tools. In turn, that should reduce development costs and time, improve maintenance and reusability, and assure commercial end users that a product has been developed following best software and cognitive engineering practice. This will allow them to assess and compare different systems and component technologies, and to choose the product that is right for them in terms of quality, price and purpose.

Putting this kind of work to practical use is the aim of another project, REWARD (REal World Applications of Robust Dialogue), in which Vocalis is also a partner. This is taking ASR technology and integrating it into practical telephone solutions for embedding in live public telephone networks. Typical applications being developed by the REWARD project include travel ticket booking, telemarketing, help desks, operator services and market research.

REWARD partners, including Vocalis, are working on prototype applications in multiple languages to enable pan-European businesses to streamline their operations across linguistically diverse markets. The project is now in its third phase, involving demonstration and deployment with target users - in the case of the UK, a telemarketing trial system is being tested by Taylor Nelson AGB

## **Speech databases and mobile applications**

One area where voice based interfaces are especially useful is in mobile applications. But noise and the variable sound quality of mobile communications make this a particularly challenging area for ASR. Vocalis is working on an EC project called, SpeechDAT-Car (Speech DATabases collected in the car for creation of voice driven tele-services), an initiative concentrating on producing speech databases, part of which is specifically geared to mobile use. SpeechDAT-Car is designed to collect speech databases to help in developing and testing recognition techniques.

Projects like SpeechDAT-Car are needed because in speech is produced differently by each speaker. Speakers of the same language have widely differing dialects, accents, and speaking rates. Their speech patterns are influenced by the physical environment, as in a car, social context, their emotional and physical state.

The SpeechDAT-Car project is being run by various teams across Europe, who will make recordings as drivers make journeys, which will be conducted under a specific set of conditions.