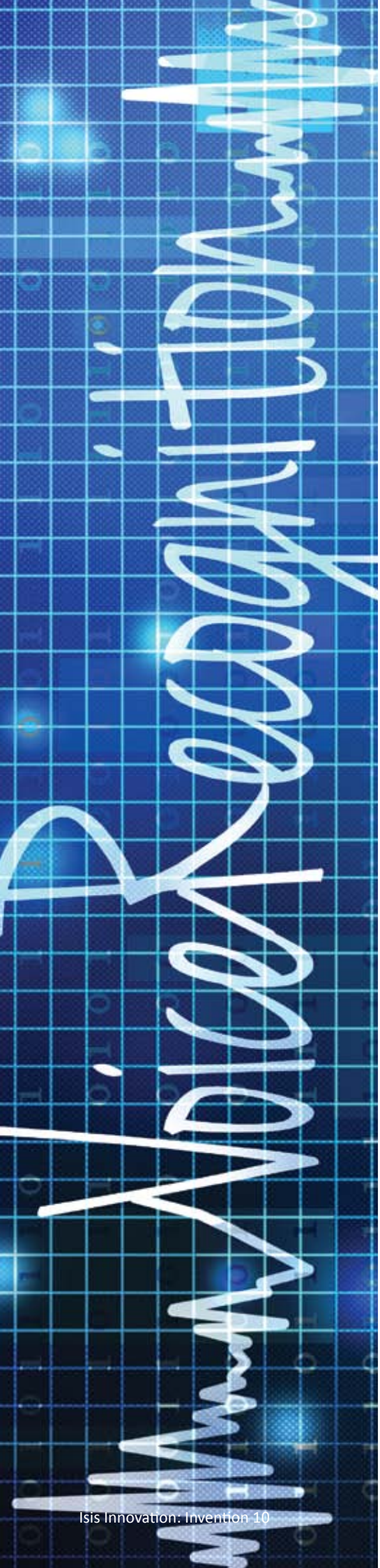
The research commercialisation office of the University of Oxford, previously called **Isis Innovation**, has been renamed **Oxford University Innovation**

All documents and other materials will be updated accordingly. In the meantime the remaining content of this Isis Innovation document is still valid.

URLs beginning www.isis-innovation.com/... are automatically redirected to our new domain, www.innovation.ox.ac.uk/...

Phone numbers and email addresses for individual members of staff are unchanged

Email : enquiries@innovation.ox.ac.uk



Speech recognition for the real world

Automated speech recognition (ASR) systems have been available for years. The challenge has always been to produce a robust system that copes well with variations in pronunciation. **Dr Fred Kemp** describes a new approach based on phonological principles

Say it again?

Is the 'a' in bath like bar or like bat? A small difference, but in reality every person pronounces every word differently, even when they repeat themselves. As a result, most Automated Speech Recognition (ASR) systems, which are generally based on statistical-modelling techniques, require extensive training from thousands of recorded speakers just to master the variation within one dialect. In contrast, Oxford's FlexSR system is based on phonological principles and therefore does not require training and can recognise a wide range of dialects and accents.

Heuristics or statistics?

Early attempts in the 1970's to use linguistic models for ASR were abandoned because they tried to use very specific algorithms to map signals to individual sounds and their variants, which required a comprehensive knowledge of the phonetic details of a particular language. Deviations from the specific mapping procedures led to the failure of these systems and the consequent dominance of statistical-modelling techniques in ASR for the last 40 years.

However, even for the current best-in-class ASR systems, high degrees of accuracy are only achieved with multi-layered and computationally-intensive models (such as Hidden Markov Models - HMMs). Such systems therefore require either state-of-the-art hardware, or in the case of mobile applications, a network connection to offload the analysis to a remote server. In addition, many systems also need to be trained against a particular voice to attain accurate recognition (although some might suggest that it is the speaker that is trained how to speak, not the software how to recognise!)

A more flexible approach

FlexSR is different. Rather than rely on statistical analysis alone, leading linguists at the University of Oxford have developed a "sparse" linguistic model of the human cognitive representation of words. This theory suggests that humans store a very basic acoustic representation of each word, accepting wide variation in the sounds themselves and recognising words by their general pattern. Adopting this approach has allowed the team to overcome the problems of earlier linguistically-based attempts, and allows FlexSR to identify words across a wide range of speakers and dialects by extracting approximate sounds and matching these patterns with its internal word list or lexicon.

By assuming there is only a small set of language and speaker independent features and allowing a wide range of variation, FlexSR even copes with the failure to detect features (either because they were not realised by a speaker or got lost in background noise). This has led to a much more accurate system, regardless of dialect, accent, non-ideal speech, or background noise. Consequently, FlexSR outperforms many existing ASR systems at individual word recognition, despite being computationally lightweight, requiring no system training and being easily adaptable to any spoken language. Currently it is implemented for English and German, including tonal languages, such as Mandarin Chinese. Tonal languages, where meaning is distinguished by changes in pitch – in a manner analogous to consonants and vowels – are extremely common in Africa, East Asia, and Central America, and often very challenging for HMM-based ASR systems.

By contrast, to train an HMM-based ASR system in a new language requires a large training set consisting of speech obtained from hundreds (or even thousands) of speakers, where the speech output has to be transcribed, segmented, and manually verified, to allow the mapping of the acoustic signal to the 'text'.

Looking to the future

The Oxford team's ground-breaking work was recently highlighted as one of the most innovative projects to be awarded a Proof of Concept grant by the European Research Council, an award that allows them to further refine and develop the FlexSR system, as well as prioritise additional languages to support.

Given the potential impact of this new approach and the broad range of applications, Isis would welcome discussions with potential partners who would be interested in integrating FlexSR into existing technologies or developing for mobile deployment.

For more information, please contact:

Dr Fred Kemp
Senior Technology Transfer Manager,
Isis Innovation
T +44 (0)1865 280919
E fred.kemp@isis.ox.ac.uk
Ref: 10377