We use cookies to ensure that we deliver the best user experience of our website. If you would like to, you can change your cookie settings at any time. Continue

Birmingham academic explores developments in voice recognition technology

Posted on Wednesday 18th January 2012

Automated voice recognition systems provide a constant source of frustration to many, especially those with strong regional accents, who often have difficulties in having their preferences recognised. Computer scientists at the University of Birmingham are currently trying to solve this problem by building a computational model of accented speech which seeks to place an individual in an 'accent space' using a small sample of his or her speech.

We spoke to **Professor Martin Russell (/staff/profiles/eece/russell-martin.aspx)**, Professor of Information Engineering from the University of Birmingham's School of Electronic, Electrical and Computer Engineering about his research into voice automated systems and his experience of training television star Richard Wilson to use voice recognition technology for **Richard's Channel 4 documentary (http://www.channel4.com/programmes/richard-wilson-on-hold/4od)**, Richard Wilson on Hold which aired on Monday 16 January 2011.

Professor Martin Russel "With just 30 seconds of speech, we can identify a speaker's region with an accuracy of 95%. We can also distinguish between different groups within a 'single accent'"

Why do voice automated systems have such a difficulty responding to strong regional accents?

You might think that speech recognition systems are based on sophisticated human expert knowledge about how human speech works, but in fact they almost all use statistical methods. For each word of the language they have an 'expectation' of what that word will sound like. They build up this expectation by analysing large quantities of real speech – this is the training phase – so the pronunciations of a particular word that they hear during training determine how they expect that word to sound when a speech recogniser is being used. If they do not experience a particular accent during training, they will have difficulty understanding a user who speaks with that accent.

How do you hope voice automated systems can be improved to overcome these difficulties?

The conventional approach to overcoming these issues is mainly by increasing the size of the training set to include more accented speech. For example, for British English you might collect several different training sets, such as Standard Southern English, Northern English, South-West English, Welsh and Scottish English. Then the speech recogniser would be able to build separate models from these different training sets and when a new user comes along, the system decides which set of models to use for the new speaker - hopefully the set that best fits his or her way of speaking and eliminates difficulty in responding to regional accents.

Although there are many variations in the ways people speak within one accent group. Even if you focus on just one city, people in different parts of the city will speak with different accents, and local people will be able to tell which part of the city they come from. Clearly the solution is not as simple as just recording more and more data. We are trying to build a system which is able to place an individual in an 'accent space' using a small sample of his or her speech. Once that individual is located in the space, we can look for other speakers who are located nearby and for whom we already have good speech models and then try to use these models to create a model for our new speaker.

We have a collection of speech called 'Accents of the British Isles' (ABI), which covers14 different accents. With just 30 seconds of speech, we can identify a new speaker's region from that set with an accuracy of approximately 95%. We can also distinguish between different groups within a 'single accent' and these advancements show promise in diluting these difficulties in the future.

What was Richard Wilson's reaction to your voice system?

I showed Richard a standard commercial speech recognition system. I did this to show him that if the system was trained properly on his voice then it would work well - and it did. Then we were able to discuss the ways in which the problems faced by a speech recogniser in an automated response system are more difficult - for example, automated systems will probably never have heard the speaker before, the speaker may be in a noisy environment, he or she is probably using a poor quality microphone which may not be held close to the mouth and the speech is probably going over the telephony system.

I believe that Richard was surprised by the performance of the commercial system that I showed him. In fact, he went away with the intention of buying one to use himself, so he must have been impressed!

© University of Birmingham 2014