A computer can pick out speech even amid cacophony

November 26 2008

Schematic diagram of SHoUT

(PhysOrg.com) -- Using a recent development in speech recognition, it is possible to search through television news programmes provided the recognition system has been trained beforehand. PhD candidate Marijn Huijbregts from the University of Twente (Netherlands) has, however, taken things even further: he has developed Spoken Document Retrieval for audio and video files that the speech recognition system has not yet been trained to deal with.

This version of speech recognition works well even if there is a great deal of unexpected background noise. Huijbregts received his doctorate from the Faculty of Electrical Engineering, Mathematics and Computer Science on 21 November.

Information can be retrieved from text very quickly using, for example, an index in a book or a search machine such as Google. However, it is much more difficult to search in audio and video files, as they do not have an easily searchable index. You can use speech recognition to simplify this process as most of the information in audio and video files

comes from speech. By recording via speech recognition, you can transform speech into text. To do this, you need a Spoken Document Retrieval (SDR) system; this makes it possible to search directly in audio and video materials, just as if you were searching in ordinary text documents. In other words, a sort of Google for audio and video.

Evening news on television

The Human Media Interaction group at the University of Twente had previously developed an SDR system for an evening television news programme. Search terms could be used to look for specific topics, the system being specially trained using newspaper texts and 20 hours of news programmes. The SDR for the evening news programme worked well because, in that situation, it was more or less known what was going to be said and there was little background noise. If you tried applying this system, without any training, to other video files, it did not perform well. Huijbregts then wondered whether he could develop a SDR system for which almost no training data would be needed, but which could nevertheless deal with unknown audio and video files satisfactorily.

SHoUT

With unknown audio and video files, it is not clear beforehand what is going to happen: who is speaking, what is being said and what sort of background noises are present. Huijbregts therefore developed an SDR system that was robust enough to deal with these unknown situations. It is called SHoUT (this acronym corresponds to the Dutch version of 'Speech Recognition Research at University of Twente'). SDR can be described as robust if it can deal with all audio and video files under all sorts of circumstances, such as background noise or if people are not speaking clearly.

SHoUT is divided up into three stages. Firstly, the system distinguishes between speech and other sounds. For example, background music is filtered out from speech. Secondly, the system identifies different speakers and gives them labels. Then finally the automatic speech recognition takes place: the system transforms speech into text. You can now search the text file for relevant topics using key words, just as Google searches through text files on Internet.

The first version of SHoUT is already available, but Huijbregts is developing it even further. SHoUT and other demonstrations of SDR systems can be found on the website of Huijbregts (wwwhome.cs.utwente.nl/~huijbreg/).

Provided by University of Twente, Netherlands