Gaelic Speech Recognition

Peter Bell, Steve Renals and Simon King, Centre for Speech Technology Research

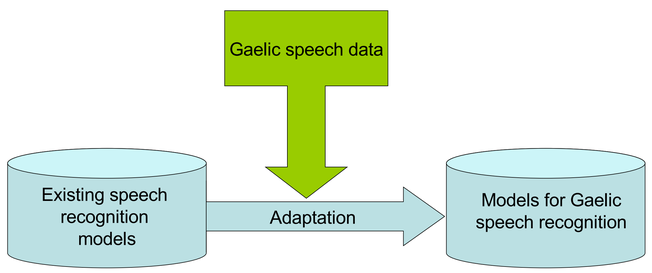
Project aims

We aim to develop a pilot speech recognition system for Scottish Gaelic. To our knowledge, there are currently no available recognisers for the language.

Building a modern speech recognition system requires large quantities of audio and text data, which can be expensive to obtain. Speech processing technology has concentrated on the most economically important languages such as English, Chinese and Spanish, for which data is plentiful, but the data requirements pose a significant barrier to the development of systems for minority languages such as Gaelic.

Building on techniques we are developing at CSTR for the EMIME project, we hope to avoid these problems by using automatic unsupervised and semi-supervised approaches that can leverage from speech recognition systems developed for different languages, using whatever audio data is available.

Cross-lingual speech recognition

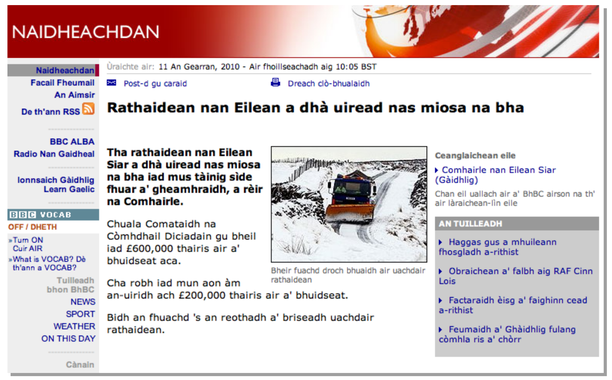


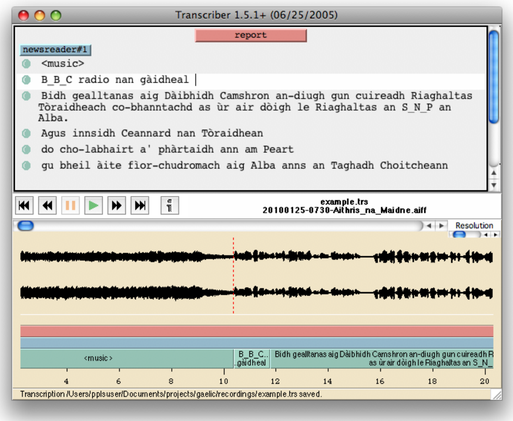
Cross-lingual speech recognition is the method of using models trained on foreign languages to aid the recognition of a target language. This figure illustrates an adaptive approach. When the training data available for the target language is limited, rather than constructing new models that may not provide adequate phonetic coverage or account for inter-speaker variability, the data could be used instead to adapt an existing well-trained model set.

What do we need?

We will work in collaboration with the Celtic & Scottish Studies department at the University of Edinburgh to obtain the resources need to build a pilot recogniser:

- around 10 hours of Gaelic speech, transcribed by fluent speakers
- a pronunciation dictionary, using a combination of linguistic expertise and automatic techniques





… and a *language model* giving prior probabilities for the words being spoken, built from text gathered from the web

Outcomes

- A pilot recognition system that could be used to facilitate the task of searching and transcribing Gaelic speech archives.
- Shareable resources to aid future projects in Gaelic speech technology.
- Development of cross-lingual adaptation algorithms that we will be of interest to the speech research community.
- Software to help the construction of speech recognition systems for other minority languages.