NS Data Release

Feasibility Study

UNIVERSITY OF CAMBRIDGE

CSaP

The Challenge

- Explore the feasibility of an **NS Data Release Scheme**, working with a candidate dataset judged to benefit both NS and academic communities.

  – follow a number of other initiatives in recent years, where Government Departments seek to derive public benefit by making data available to researchers (eg Department of Health and HM Revenue & Customs);

  – track the work of the **Administrative Data Taskforce** (ADT), established to propose new mechanisms and collaborative agreements to enable and promote the wider use of administrative data for research and policy purposes; it reports to Ministers in December 2012.

The Candidate Dataset

- NS stakeholders have nominated, as the candidate dataset for this study, a **speech corpus** comprising a database of speech audio files, text transcripts and lexicon. This could be used by researchers to build linguistic and acoustic models to support research into speech technologies.
- The corpus comprises:
  - c 400 hours of unclassified data collected commercially,
  - c 500 hours of classified data collected through telephone interception authorised by the Regulation of Investigatory Powers Act (RIPA).

Research Benefits

Opportunities

– substantial National Security benefits would accrue if speech technology tools could be developed and applied to operational data; this opportunity is significantly increased by involving one or more high-quality research teams in UK universities and by using realistic data.

Research Challenges

– Leading academics say there are non-trivial research challenges that would benefit from using this Speech Corpus. This provides the chance to tackle speech in "natural situations" where they can address:

- realistic acoustic problems (e.g. interpreting words spoken in an echoing house)
- linguistic problems, including speech patterns, vocabulary and accents ;
- understand how signals work, overcoming degradation created as sounds are coded, transmitted and decoded across different communication channels.

Infrastructure

There are a range of logistical options for administering the classified component of the Speech Corpus:

- Government premises
  - £250,000 for establishing a stand-alone system;
  - higher level of security assurance
  - a detrimental impact on the working-practices and activities of the academics concerned (who have to conduct research away from their academic institution).

- Academic Premises
  - ease of use for on-site researchers;
  - security operating procedures have to be developed for the site;
  - similar financial costs for secure stand-alone, security-accredited system … multiplied if more than one university involved.

Data Owner Concerns

Legal/Ethical

– The data-owner is obliged by statute to control the disclosure of information that it has obtained; this is only allowed in so far as it is deemed necessary for the proper discharge of its functions;

  - See Section 15 (2) of the Regulation of Investigatory Powers Act (2000)

Operational

– Senior managers with responsibility for the operational acquisition and use of the classified component of the Speech Corpus data have recognised the potential benefit to their mission in making the data available for research.

Security

– there are risks, but these can be mitigated through the careful selection and editing of data to eliminate any significant sensitivities and by the application of appropriate security procedures - notably accreditation of any system holding the data and security clearance of personnel accessing the data.

Provisional Opinion of Deputy SIRO

On the unclassified component of the Speech Corpus

The data-owner should be free to transfer this material to a speech repository, on the understanding that participants understood the data was going to be used for research.

Provisional Opinion of Deputy SIRO

On the classified component of the Speech Corpus

- ***Assuming that the business case and legal and ethical views are in order, my recommendation to SIRO and the Information Board on this request would be that the information risks are manageable provided:***
  - ***we have assurance that legally privileged or sensitive data and explicit identifiers are removed (with other protective measures being taken as appropriate);***
  - ***the stand-alone system will be accredited to SECRET standard for use by individuals cleared to SC;***
  - ***we would expect the contractual arrangements to include the commitment to confidentiality, security procedures and disclosure controls.***

Given the unusual, ground-breaking nature of this request, the SIRO should brief the authority's Information Board and, then, ensure that the Interception Commissioner is satisfied with our approach and mitigations.

Conclusions

- **We have demonstrated that in principle sensitive NS data can be released for research, where this supports the statutory function of the data-owners**.
- The release of the unclassified data-set to a research repository is also endorsed – a quick win.
- The authority clearly takes its responsibilities under the law seriously, with the gravity of the question reflected in the need to consult the Interception Commissioner.
  - A number of careful steps would have to be taken to design effective arrangements.
- Ultimately, the expenditure of NS resources on this project would have to be weighed against competing priorities.

Are there opportunities available from the findings of the Administrative Data Taskforce?

UNIVERSITY OF CAMBRIDGE

CSaP