

LANGUAGE RESOURCES OF THE (NEAR) FUTURE

Nick Campbell
Trinity College Dublin

SPEECH TECHNOLOGY

- Current speech technology maps between speech and text, but it can't model the information flow in real interactive speech
 - propositional content + discourse controls that simultaneously signal cognitive attentional, intentional, and emotional states.
- corpora for future speech technology must be multi-modal, multicultural, and multilingual
 - for research into interpersonal communication strategies, and their related bodily movements, and the characteristics of interactive speech for advanced speech interfaces

SPEECH CORPORA

- Many multimodal corpora have been reported at Interspeech and LREC, but there is still limited understanding of the best ways to gather representative samples of interactive speech, or of how to control the necessary variation and range of expressivity without rendering the resulting conversational interaction unnatural or artificial in some way.

(INTERACTION STYLES)

- from read speech to spontaneous, from isolated words to sentences, to telephone conversations and meeting dialogues ...
- the focus has been on the content of the speech rather than on the style of interaction ...
- future collections should concentrate on controlling the interpersonal aspects of the speech, not the content of the speech itself

SEMANTIC ANNOTATION

- The processing of 'meaning' in speech remains a significant challenge.
- An ISO standard currently being considered for dialogue act annotation (ISO/TC37/SC 4 N442 rev05) proposes eleven (11) classes of discourse acts, including *task-related*
- Task-related discourse is lexically very complex, and will continue to pose a (possibly infinite) problem for speech technology

(LEXICAL COMPLEXITY)

- however, the remaining ten (10) major categories of spoken interaction all feature limited lexical content but they are characterised by complex prosodic variability.
- so perhaps the most immediate task for data collection (in the short-term future of speech processing research) is how to address the problem of the sparse coverage of these ten discourse categories.

MULTICULTURALITY

- technology should serve all aspects of society, young and old, technical and professional, educated and naive, all allowed equal access ...
- but we don't yet know the needs, or even the identity, of these different classes of society ...
- balancing the sociological aspects of data collection will become a major concern

MULTILINGUALITY

- *Universal Linguistic Rights* require the provision of language services for all people in their own mother tongue
- The provision of governmental, commercial, and individual services in each person's own language is a right that our technology can facilitate.

(REPRESENTATION COSTS)

- ... but the collection of representative resources for the less-well represented languages of the world remains a major challenge!
- Tools and specifications ('blarks') exist for such collections, but allocating funding to cover the less-well represented languages of the world must remain high priority.

MULTIMODALITY

- the focus of speech corpora has been on the speech itself but there is a growing awareness of the multimodal nature of interactive conversational speech
- the use of video, as well as audio, is now almost compulsory ...
- and 360-degree video and motion capture is also becoming common

(DATA PROLIFERATION)

- ... so the amount of data in a corpus is growing at ever-increasing rates.
- This poses difficulties both for the annotation and storage of the material as well as for its distribution ...
- particularly when data collections span several days, months, or even years!

CROWD SOURCING

- we find an increasingly social aspect to the internet, and see the emergence of crowd sourcing for a variety of tasks ...
- The CS annotation of speech data will provide a rapid and reliable source of knowledge, incorporating the views of a wide range of people from across the world ...
- but provision of easy access to such large amounts of material remains a challenge !

DATA DISTRIBUTION

- Distribution of resources will continue to be difficult, and increasingly more so as the volume of data increases and includes multi-source video and biometric information ...
- but proper structuring of metadata will allow fast access to specific subsets of the data, and by selectively varying the compression rates and podcasting the material, wider and more rapid methods of distribution will be achieved.

RIGHTS MANAGEMENT

- The collection of natural unstructured conversational speech material necessarily concerns the possibility of infringing personal rights when distributing such materials.
- The legal aspects of ownership of speech material, and the inadvertent information pose important problems that remain to be addressed.

EXPRESSIVITY

- Speech is not a simple information source but signals under-specified *propositional* and *interpersonal* content simultaneously.
- physics has mastered the modelling of complex quantal phenomena without requiring any single interpretation to be dominant in absolute terms
- we might benefit from adopting similar models for the processing of *meaning*, and for the interpretation of spoken communication

SUMMARY

- speech is complex
- corpora are proliferating
- multifaceted aspects introduce complexity
- but these are not insurmountable problems ...

CONCLUSION

- thank you for listening :-)

