Multi-Modal Data: Speech

Project Seminar: Unlocking the Secrets of the Past

Daan R. Henselmans

February 23rd, 2011

Samuel Gustman, Dagobert Soergel, Douglas Oard, William Byrne, Michael Picheny, Bhuvana Ramabhadran, Douglas Greenberg. ”Supporting access to large digital oral history archives”. *Proceedings of the Joint Conference on Digital Libraries*. 2002.

Overview

- Shoah Visual History Foundation
- System Architecture
  - Issues and Proposed Solutions
- Automatic Speech Recognition
- Text Processing
- Retrieval Algorithms
- User Interfaces

Shoah Visual History Foundation

- 1994 initiative of Steven Spielberg
- Performing four tasks:
  - Collecting and preserving survivor and witness testimony of the Holocaust
  - Cataloging those testimonies so they could be made available
  - Disseminating the testimonies for educational purposes to fight intolerance
  - Enable others to, or perhaps have the VHF itself collect testimonies of other atrocities and historical events

Shoah Visual History Foundation

- Today:
  - 52,000 testimonies
  - 32 languages, representing 56 countries
  - 116,000 hours of video
  - 180 terabyte digital library in MPEG
- Clip boundaries, summaries and descriptors
- Eight documentaries, two CDROMS, several museum exhibits, and one book
- Collection techniques, digitization workflow, and support for human cataloguing freely available

System Architecture

- Video Data
  - Beta-SP digitalized to MPEG
  - Average duration of testimony just over two hours
  - Cut into distinct ”clips” of 3.5 minutes on average
- Metadata
  - Interview Details
    - Pre-Interview Questionnaire (PIQ)
  - Release Status
  - Interviewer Data
  - Descriptors, properties, and summary of each clip

System Architecture: Issues

- Manual Cataloguing
  - Three-sentence summary of each clip
  - Links to appropriate thesaurus descriptors
  - Requires **about 15 hours** per hour of video
- Full-description clip-level cataloguing for 116,000 hours of video would cost over $150 million
- Time consumed mostly by establishing clip boundaries and writing clip summaries
- Emotional content lost in summaries

System Architecture: Proposals

- Real-time cataloguing system
  - Automatic determination of clip boundaries
  - Automate portions of cataloguing process
- Search function
  - Whole-testimony level:
    - PIQ Data
  - Within-testimony level:
    - Automatic Speech Recognition (ASR)
    - Descriptors assigned by automatic summarizers
- Can be used directly, or to assist human cataloguing

Automatic Speech Recognition

- Two Steps:
  - Recognition of phonemes
  - Derivation of terms
- Driven by Statistical Models from Training Data
- Ngrams map to word classes (names, places)
- VHF Thesaurus used to obtain word-to-class mappings
- May be dependent on language community

Automatic Speech Recognition

- Requires improvements to ASR
  - Acoustic Segmentation: dividing the acoustic signal into categories of speech (emotional, speech in different languages, etc.)
  - Rapidly adjusting acoustic model to speaker
  - Task-dependent functions geared toward retrieval
    - Giving higher weights to words that are important for searching and automatic classification
- Obtain names from a large list pertinent to the domain
- Goal: Provide sufficient word and phrase information for further text processing

Text Processing

- Determination of Clip Boundaries
  - Combine acoustic segmentation with semantic models

Text Processing

- Assignment of Descriptors
  - Clip level: classifier scans testimony and assigns a descriptor if enough evidence can be found
  - Testimony level: derived from the set of clip-level descriptors by consolidation and abstraction
- Summaries formed as sets of descriptors
- Fluent summaries may not be possible
- Assign degrees of confidence
  - Human editor can focus on pieces the machine could not do well

Retrieval Algorithms

- Many types of evidence:
  - Phonemes
  - Terms in testimonies
  - Time proximity
  - Descriptors in thesaurus
    - Different scopes for different categories
      - i.e. place names have a big scope, while activities have one of a few minutes
- Possibilities for retrieval based on any of these types used singly or in combination
- All must be extended to cross-language searching

User Interfaces

- Query frame with categories of criteria
  - Assistance with finding the right descriptors
    - Mapping free-text entry vocabulary to nominate thesaurus terms
    - Browsable thesaurus hierarchy
- Fast access to audio or video
  - Very important if no or limited surrogates are available
- Assistance to users in defining and grouping clips
- Presenting a time line, a map, and images

Conclusion

- Shoah Visual History Foundation: a large digital oral history archive
- Browsable testimonies and clips
- Providing specific access to data is problematic
- Proposal: a research agenda covering issues in speech recognition, classification, retrieval and more
- Much work remains to be done!

Resources

- Samuel Gustman, Dagobert Soergel, Douglas Oard, William Byrne, Michael Picheny, Bhuvana Ramabhadran, Douglas Greenberg. ”Supporting access to large digital oral history archives”. *Proceedings of the Joint Conference on Digital Libraries*. 2002.
- Official website: http://college.usc.edu/vhi/