University of Twente
The Netherlands

Spoken Document Retrieval

Franciska de Jong

University of Twente
Department of Computer Science
contact: fdejong@cs.utwente.nl

Paris - LangTech - 24 November 2003

What is Spoken Document Retrieval?

- *functionality*: support access to fragment of spoken audio (in radio/video archives, meeting recordings, telephone taps, etc.) via automatic metadata generation
- *how*: combination of technologies: coding, storage, speech recognition (speech-to-text transcription), indexation, searching
- *crucial*: automatic generation of time-coded index
- *where*: application to be incorporated in workflow environment (production, postproduction, archiving)

Why are timecoded indexes useful?

- massive A/V data repositories turned into content with value *at fragment-level*
- domain specific processing may generate exclusive data sets
- Cf. Kenneth Church at Eurospeech '03:

  If petabytes are coming,

  - demand for data storage will not keep up with supply
  - search will become a killer application

Speech Transcription and Indexation

Transcription:

- conversion of speech into text (series of words)

Indexation:

- application of full text indexing to transcribed segments
- feed of time-coded index into a specific meta-data field
- advanced: summarization, clustering, …

Transcription vs. Dictation

Automatic transcription of broadcasted speech can not be performed using a standard dictation system:

- Segments of different acoustic nature (studio quality, noise, telephone, music, overlapping speech)
- Segments of different linguistic nature (read speech, prepared speech, spontaneous speech)
- Wide variety of speakers (news anchors, reporters, politicians, common people, dialects, non-native)
- Wide range of topics, topics change over time (requires language modeling, based on huge qts of textual data)

Speech recognition: State-of-the-art

Contemporary broadcast news transcription:

- 20% word error rate (in international benchmarks)
- Word error rate highly dependent on speaker and speaking style (ranging from 1-2% to over 50%)
- Comparable results on several languages (English, French, Mandarin, German, Italian, Spanish, Portuguese...)
- Application development slow for several languages (including those for smaller markets, e.g. Dutch)

Audio Partitioning

Preprocessing of audio

- Remove non-speech segments (music, noise)
- Improved speech recognition by speaker/condition adaptation
  - Identify speaker turns and speakers (relative or absolute)
  - Use acoustic models specific to the condition (narrowband/wideband, male/female)

Current retrieval performance

- Recognition error rates for content words are better than for function words
- Estimated retrieval performance:

  average precision > 50%
- Conclusion: sufficient accuracy for audio fragment retrieval is feasible

University of Twente
The Netherlands

What is out there to be disclosed?

- broadcast material (news and other)
- governmental proceedings (parliamentary sessions, court recordings, commissons)
- oral history narratives (retrospective interviews)
- presentations (speeches, lectures, readings)
- interactive meetings (business, medical teams, conventions, etc.)
- recorded telephone conversations (private, business, teleconferences)
- cultural heritage

University of Twente
The Netherlands

Examples of digitised A/V collections

Annotated (not via ASR!) and accessible via the web

- News and cultural programming
  - RAI Radio (Italy)
  - BBCi (UK)
  - National Public Radio (US)
- Other
  - US Supreme Court Sessions
  - …

Other

- 30 channels of Dutch radio, video and web broadcasts (since 2001; including parliamentary sessions)
- sessions of Yuguslavia Court (The Hague)
- the MALACH collection
- Apollo Mission Archives

University of Twente
The Netherlands

The MALACH Project

- 52,000 interviews with Holocaust survivors
  - 116,000 hours (180 TB MPEG-1)
  - 32 languages, recorded in 67 countries
- Full description cataloging: 4,000 interviews
  - Manual segmentation
  - Thesaurus descriptors (14,000-term polyhierarchy)
  - Structured segment summaries
- "Rapid" cataloging (120 person-years)
  - Time-tagged thesaurus descriptors

Running SDR products/systems

ASR Applications for broadcast news

- Virage (English television; commercial)
- Speechbot (English; web search on radio)
- DRUID - Dutch speech recognition on radio Journal (experimental)
- ..... (many more)

SpeechBot

University of Twente
The Netherlands

Simple Search | Power Search | Help | FAQ | About SpeechBot | Feedback

Search for: organic food [Search]

Topics: All Topics Dates: All dates

Tip: *Try searching a particular topic instead of "**All Topics**"*

Search Result: **200 matches** for your query Sort results by: Relevance

	Website	Date	Extract from Transcript (*Transcripts based on speech recognition are not exact*)
PLAY extract	The Diane Rehm Show	Mar 14, 2000	...new standards for growing and processing **organic food** the proposal incorporates recommendations that consumer groups and **organic** farmers... Show me more
PLAY extract	PBS Online NewsHour	Dec 21, 2000	...advantage of the exploding demand for **organic** products glickman also said the standards would make things a lot clearer for consumers 1 northern virginia **organic food** shopper... Show me more
PLAY	Public Interest	Aug 28, 2000	...get us on the right direction and but you move for a new now we can support farmers markets weekend the point

20 extracts from **The Diane Rehm Show - Mar 14, 2000** match your search:
organic food

You are here
extract 1 of 20

NEXT EXTRACT

PLAY extract

...waga in new and marching tune and diane ream you as to prevent it the agriculture is proposed new standards for growing and processing **organic food** the proposal incorporates recommendations that consumer groups and **organic** farmers but some say that standard to comply with popular opinion rather than scientific research joining me to discuss **organic food** standards can claim they're against chief marketing standards in and the state years for the u. s. department of agriculture tell a d. on stand and ...

Display 30 seconds of transcript

Extracts from this transcript in order of relevance:

standards for growing and processing organic food the proposal incorporates

PLAY 51 min

The Diane Rehm Show - Mar 14, 2000

Visit this website
Search all transcripts from this website

University of Twente
The Netherlands

More than transcription

Transcription: which words were spoken and in which order?

But also other important types of metadata needed:

- Who said what? (*speaker identification*)
- Which topics were addressed? (*topic dectetion and classification*)
- How did topics develop over time? (*topic tracking*)
- What events (e.g. meeting acts) occurred? (*information extraction based on domain models*)
- What other modalities (gesture, lip movement) can be used for disclosure? (*fusion technology*)
- What related material is available? (*linking* and *multisource content browsing*)

And added value for non-time coded files (*insertion of time stamps in human generated transcripts)*

Research issues (beyond ASR)

- retrospective digitisation
- deteriorating quality for older analog repositories
- development of dedicated browsers
- copyright/privacy
- evaluation methodology
- annotation standards
- acquisition of training collections

Some example projects/communities

- Informedia (CMU, US)
- EU 4/5th Framework: Olive, ECHO
- Current and upcoming EU projects:
  - Indico (lecture recordings)
  - M4 (meeting recordings)
  - AMI (meetings recordings; inc. multimodal aspects)
  - .....
- NSF/DE LOS working group on spoken word audio archives
  - report at http://www.dcs.shef.ac.uk/spandh/projects/swag/
- DARPA EARS program (rich transcription)