automated speech and audio analysis for semantic access to multimedia

Franciska de Jong, Roeland Ordelman, Marijn Huijbregts
Human Media Interaction, University of Twente
The Netherlands

overview

1. variety of multimedia content and semantic annotation
2. automatic metadata generation by exploiting linguistic content:
    1. collateral data
    2. audio analysis:
        1. speech recognition
        2. audio classification
3. summary

content is king ...

- news (professionals & non-professionals)
- historical material (retrospective digitization)
  - interviews (oral history)
- meetings (governmental & corporate)
- presentations, lectures
- private material (lifelogs?)

but metadata rules!

- networked electronic media and increasing size require automation of content-based extraction
- use linguistic content in multimedia archives: boost the accessibility
- Two examples:
  - 1995: subtitling for broadcast news retrieval
  - TRECVID: best performing systems exploit speech transcripts

semantic gap

- bridge gap between user needs and content features
- traditional approach: manual annotation + controlled vocabulary index terms
- enhance traditional approach: automatic metadata generation
- challenge: combine various types of metadata and exploit the added value

2. automatic metadata generation by exploiting linguistic content

manual annotation

use collateral data

audio classification

speech recognition

use audio analysis

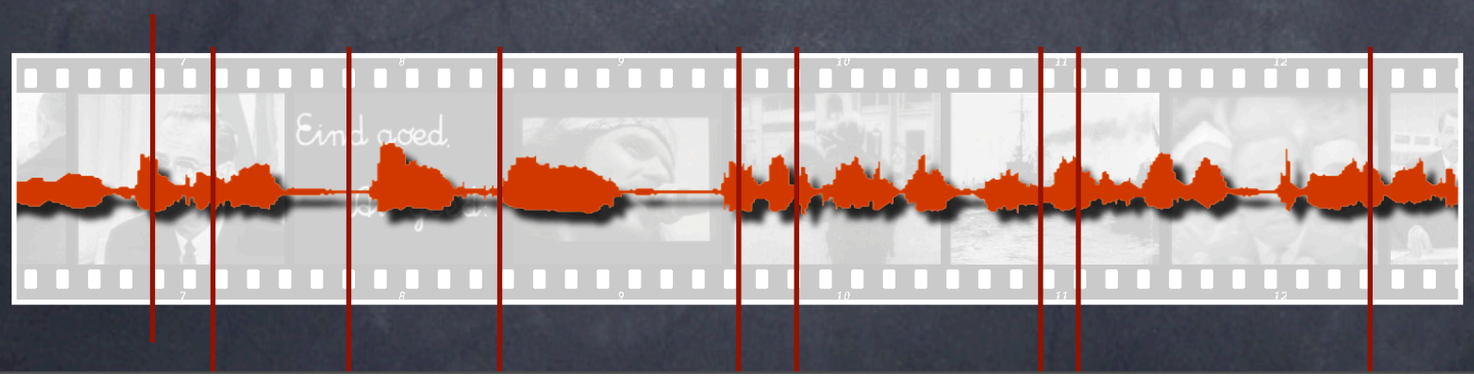
metadata for searching

exploiting collateral data

- broadcasts:
  - subtitling information for hearing impaired (time-codes included; Dutch subtitling has topic-boundaries)
  - teleprompter files, recording scripts
- other:
  - meeting minutes, notes (e.g., soccer matches, interviewer notes, presentation notes)

exploiting collateral data

- collateral data: locate documents
- searching within documents: alignment of text to multimedia document
- label text with time-codes: speech recognition techniques (alignment)



alignment

- speech recognition training: forced alignment
- alignment: find most optimal distribution given audio + the words that are spoken
- caveats: large chunks, missing parts, model mismatch (e.g., low audio quality, old-fashioned speech)
- solutions: model adaptation, two-pass strategy

2.1 exploiting linguistic content: collateral data

Continuous Access To Cultural Heritage

HMI
Human Media Interaction

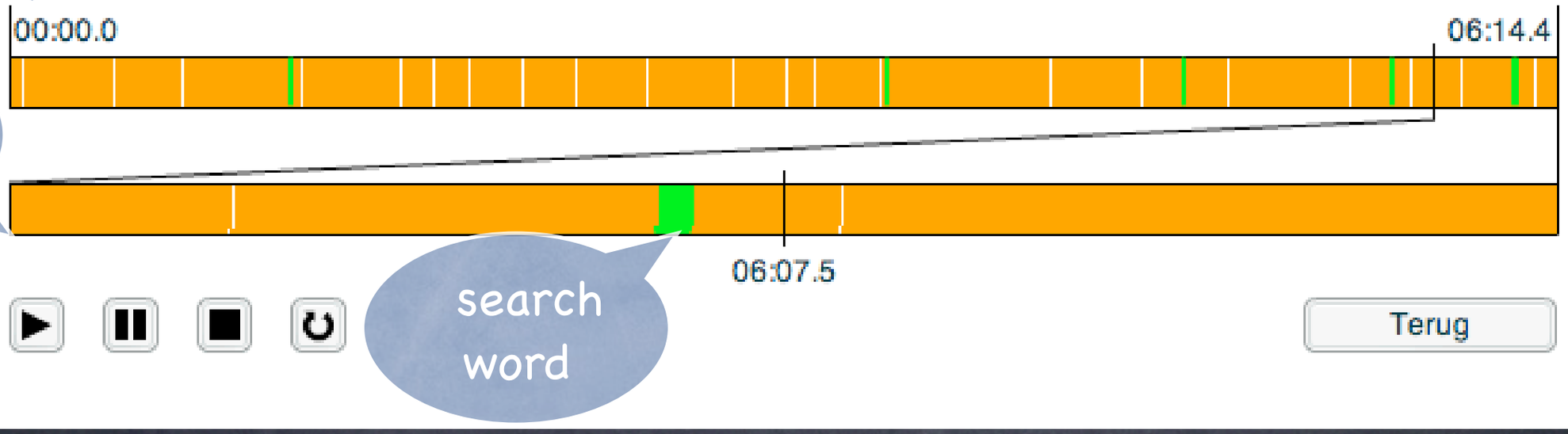
add a picture based on speech content

subtitling following speech

bevrijding van uw leven letterlijk en geestelijk

complete speech

relevant segment



search word

Browsing and searching WWII speeches of Dutch Queen Wilhelmina

cross-media mining

- not only synchronize audiovisual material that approximates the speech
- any kind of textual resource that is accessible: , open source titles and proprietary data
  - trusted webpages
  - newspaper articles
- shift focus from indexing individual documents to indexing multiple multimedia databases

cross-media news browsing

CV of principal person

on topic governmental publication

on topic newspaper article

news segment (indexed on subtitles or speech)

related news segment

Geef het onderwerp waar u op wilt zoeken:

Zoekterm(en) hypotheekrenteaftrek

Informatie uit: Teletext Spraakherkenner

Sortering Kranten
Datum Krant
Score KrantSegment (1)
Score JournaalSegment (2)

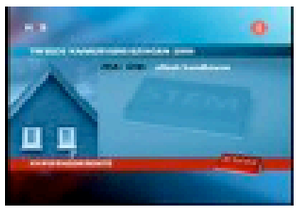
(1) hoe scoort de zoekterm in een krantenartikel
(2) hoe scoort de tekst van een journaalsegment in een krantenartikel

Zoeken in Journaal

[NB: Aantal krantenartikelen in database: **1419408**]

[Klik op thumbnail om journaal te bekijken]

ZOEKRESULTATEN



Journaal maandag 13 november 2006 (00:00:41)

- Score: 11.

...De partijen willen dat er meer voor starters gebouwd wordt de bedrijven waar wij nu over Gunco bonus voor starters hoeven geen overdrachtsbelasting te betalen. Telia wil een stakers fonds voor goedkope leningen en de VVD denkt dat meer bouwen. Oplossing is dan **de hypotheekrente aftrek** twee jaar WW junior Freek Vossius handhaven de bedrijven daar bij de buren voor de nieuwe gevallen beperken tot twee en_ veertig procent nu strenge vijftig procent van de rente **aftrek**baar ook **GroenLinks wil de aftrek** beperken voor alleen de lagere en middeninkomens en de SP legt de grens waarvoor bij hypotheken **tot drie honderd vijftig duizend euro. ...** »

Krant	Datum	Score Krant	Score Journaal		Artikel titel
de Volkskrant	6 jun **2003**	*8.897*	*42.79*	✓	Huizenprijzen »
AD www.ad.nl	28 mrt **2002**	*10.93*	*38.42*	✓	Heilig huisje »
AD www.ad.nl	2 mrt **2002**	*6.783*	*39.52*	✓	PvdA: Maximale **aftrek** jonge huizenkoper »
AD www.ad.nl	12 dec **2001**	*12.24*	*49.70*	✓	PvdA wil jonge huizenkopers bevoordelen »
Trouw	26 aug **2000**	*10.50*	*37.96*	✓	Nederlands uniek, maar hoelang nog? »

speech indexing

- speech recognition: full text annotations
- spoken document retrieval in the American-English broadcast news (BN) domain was declared "solved" with the NIST-sponsored TREC SDR track in 2000

speech indexing:main steps

- train acoustic models using speech that resembles speech in task domain
- train language models using text data that resembles speech & word usage in task domain

- de-multiplex, audio-conversion
- segmentation: speech/non-speech, speaker-diarization
- recognition (multi-pass: speaker/LM-adaptation)

- post-process speech transcripts (format, rich text)
- index & search

challenges

- rough baseline: at least 50% speech recognition accuracy required for IR
- speech type: planned/formal - spontaneous/ informal - cross-talk
- speaker characteristics: non-native, dialect
- audio quality (recording conditions, acoustic environment)
- availability of training data
- vocabulary (OOV - QOV)

audio classification

- speech (words)
- structure (silence, music/jingles, speaker)
- speaker:
  - identity (gender, native, dialect, age)
  - emotion

summary

- King-size multimedia content repositories
- require content-based extraction techniques
- exploit readily available collateral data and audio analysis techniques
- (multiple) linguistic annotations allow for cross-media browsing
- challenge: how to combine various types of metadata

PS: has this been recorded?