Dublin City University

School of Computing

CA4009: Search Technologies

Section 9: Multimedia Search

Gareth Jones

November 2016

Introduction

Much information is contained in media other than electronic text:

- spoken content - radio broadcasts, recordings of meetings and lectures
- printed items - legacy texts, e.g. books, newspapers, corporate documents
- static images - photos, cartoons
- video - sports, films
- music (recordings), music (notation)

Many sources involve a combination of different media, e.g. video, speech and music in a movie.

Introduction

Many of the techniques developed for search of electronic text data can be applied to information held in other media.

However, a number of features of these other media introduce new issues which must be addressed to deliver useful search applications.

The key additional features which must be taken into account are:

- the need to process the content to recognise what is actually in the data, so that they can be *indexed* for search and retrieval.
- how to specify a search request, e.g. how should a query for an image be entered?
- how to interact with or browse the retrieved results, e.g. how to recognise relevant content, how to audition the contents of an audio or video file.

Specifying Requests

While appearing appearing simple or obvious, multimedia search requests are often complex to interpret and perform.

Examples of some basic multimedia search requests:

- Find me movies containing car chases.
- Find me goals in soccer matches.

Real requests are likely to be more specific:

- Find me tries by Brian O'Driscoll in his last 6 nations series.
- Find explanations of database normalisation in online lectures.

Specifying Requests

Within a specific video. more detailed browsing requests might include:

- Show me the car chase in this movie
- Show me the goals in this soccor match.
- Personalised summary - Show me a summary of this F1 race focusing on Lewis Hamilton.
- Query by example - Is the person in this picture in this movie?

Content Recognition

For speech data:

- we can user automatic speech recognition, but this can be difficult - see later.
- increasingly we can user crowdsourcing - where human volunteers transcribe the content in return for a small payment (how could we check the accuracy?), but much spoken data will be confidential, so do we want to let other people listen to the content to transcribe it?

For printed content:

- Use optical character recognition (OCR) - again much content can be challenging.
- Again we can use crowdsourcing.

Content Recognition

Images:

- How should we analyse in an image?
- What should we try to identify? What features will be useful?

Videos:

The same as images, but also

- Multiple frames showing very similar scenes - can make analysis more reliable.
- Temporal dimension - find events.

Again, automatic analysis or crowd labelling.

Content Recognition

Music:

- What should we try to extract?
  - what will be useful to support users in search?
  - melody, chord progressions, lyrics, rhythm, structure of the composition
  - instruments playing
- Automatic audio feature extraction - this is difficult
- Analysis of printed notated music.
- How can we use computer representation of music, e.g. MIDI files.

Interacting with Retrieved Results

- How should we represent content to determine relevance?
  - images? keyframes from video?
  - speech/music - listening to whole item too time consuming.
  - audio snippets - how should these be created? excerpts from speech transcripts?
- How should we access the content in a relevant item to satisfy the user's information need?

Scope of Multimedia Search

Multimedia Search (or Multimedia Information Retrieval (MIR) is a very active research area which seeks to:

- better understand user needs and their cognitive abilities and preferences for media interaction,
- develop technologies to create effective multimedia search systems.

Solutions described here should not be regarded as definitive; new ideas are being proposed, developed and evaluated all the time.

Topics for Multimedia Search

- Retrieval of Spoken Content
  - introduction to the principles of speech recognition
- Content-Based Retrieval of Visual Media
  - features for visual indexing
- MIR systems
  - search
  - browsing

Retrieval of Spoken Content

From a search perspective the retrieval of spoken documents is the most straightforward extension from text retrieval.

Retrieval can be based on:

- text requests to search spoken data,
- spoken requests to search text data,
- spoken requests to search spoken data.

content.

Retrieval of Spoken Content

Most obvious challenge of spoken content retrieval is that the contents of spoken documents are not immediately available for indexing. i.e. we don't know what is being said!

An ideal solution to this problem would be to generate a perfect transcription of the spoken data, and then treat it as text data and apply standard text information retrieval methods.

But, perfect transcriptions are frequently not available:

- perfect automated transcription is not possible;
- full professional manual transcription is generally usually uneconomic, e.g. business meetings;
- crowdsourcing may be available for non-confidential content, e.g. radio and TV content.

Retrieval of Spoken Content

Automatic speech recognition (ASR) systems can be used to generate imperfect index information for spoken content.

Index errors arising from errors in speech recognition will reduce the performance of a retrieval system (how?), but in practice retrieval is usually good enough to be useful (why might this be?).

The level of reduction in retrieval effectiveness can be explored using quantitative experiments to measure variations in metrics such as the *precision* and *recall*.

For example, an experiment to investigate the change in MAP which arises from the users of a errorful transcripts created using an ASR system compared to a perfect manual transcript?

Retrieval of Spoken Content

While the usefulness of the search system to a potential user can be indicated by the results of quantitative experiments, it is also important to explore the user's response to the system via user evaluation.

What does reduction in MAP mean in terms of the user experience?

Is a system which has a 10% reduction in MAP acceptable to the user? Might 30% reduction in MAP be acceptable?

Speech Recognition: An Introduction

Human processing of spoken language is highly sophisticated, and simulating this using a computer requires algorithmically complex software.

Speech recognition can be thought of as a *decoding* task.

– specifically the decoding of air pressure wave signals into a written word signal.

In a computer, the speech signal is captured using a microphone and an audio amplifier.

This signal is then sampled using an analogue-to-digital converter.

The resulting digital signal can then be processed by the computer.

Challenges of ASR

ASR is challenging for a wide range of reasons including:

- speech variability
- speaker variability
- acoustic ambiguity
- continuous speech
- context-dependency

Speech Variability

Each utterance of a word is unique. Utterances may vary in many ways including:

- speed of delivery.
- the pitch of the voice.
- the pattern of stressing.
- the volume.
- background noise conditions.

It might be considered remarkable that human listeners have the perception that the same word is spoken on each occasion!

Speaker Variability

A further significant challenge for ASR systems is that the speech patterns of different people show significant variations.

These variations include:

- variations in pitch,
- vocal qualities,
- linguistic expression,

even when people are trying to perform similar tasks.

In order to produce speaker *independent* systems that recognise the speech of *any* individual, acoustic models of speech must be rich enough to capture all of these variations.

Speaker Variability

True speaker independence is not currently possible:

- American and British English systems use different acoustic models,
- can we buy an Irish English speech recognition system?
- non-native speakers represent a significant challenge.

Speaker “independent” models are trained using speech examples from many people from a wide variety areas and of different ages.

If sufficient training data is available speaker *dependent* systems will tend to perform better.

Some systems adopt a technique called *speaker adaptation* where speaker independent models are adapted to a certain speaker over time using a small amount of spoken adaptation data.

Acoustic Ambiguity

Some words are acoustically ambiguous.

For example,

to, *too* and *two*

are homophones which sound the same.

Also, some words only have a very small acoustic distinction.

For example,

bee and *pea*,

while not true homophones, can be highly confusable, especially when the initial consonant is not well articulated.

Isolated vs Continuous Speech

Early speech recognition systems were designed to recognise only individual words with short silences before and after each word. This is referred to as *isolated* or *discrete* speech recognition.

Current systems allow *continuous* speech recognition, i.e. recognition of flowing speech without silent gaps between words.

In addition to not knowing what the words are:

- the recogniser doesn't know where the word boundaries are,
- or even how many words there are!

Styles of speaking in fluent speech make the task still more difficult.

- For example, in the phrase *you need to know* the *d-t* can be merged giving a phrase more like *you nee'to know*.

Context-Dependency

All words can be broken down into a small set of constituent sounds, referred to as the phones of the language.

English has about 45 distinct phones.

For example, `speech` is composed of the phones `s p iy ch`

Each phone can be produced a number of ways depending on its context, leading to slight variations in its sound.

Speech Recognition Systems

Speech recognition systems typically comprise two fundamental components:

- Acoustic models - statistical models of all the possible speech sounds of the language that is being spoken.
- A language model - a statistical model of the expected word (or other units to be recognised) sequences of the language.

These are combined in a process referred to as *decoding* to generate the most likely output sequence, i.e. the most likely sequence of word spoken.

Acoustic Pattern Matching

A separate acoustic model is built for each acoustic unit, usually a word or subword phone unit.

The most common current approach to acoustic pattern matching is statistical modelling using hidden Markov models (HMMs).

In very simple recognition tasks with a vocabulary of a few words no language model may be used with recognition relying only on the acoustic model.

The word model with the highest matching score when compared to the acoustic input is the recogniser output.

- which doesn't necessarily mean that it's the right word!.
- even an apparently simple ASR task can be difficult - see above for why.

Word vs Subword Recognition

Current speech transcription systems typically have a vocabulary of around 80,000 words.

Attempting to recognise 80,000 word models in parallel is not computationally possible.

Also there are significant issues in the availability of training data for this many words. .

The required recognition task can be dramatically reduced by performing recognition at the subword phone level.

Recognised phones are mapped to words using a phonetic dictionary.

Word vs Subword Recognition

However, phone recognition is itself hard due to the acoustic ambiguity and shortness of phones.

Typically best performance will be about 70% correct, but it is often worse.

A simple phone recognised output will be a noisy string in which it is going to be difficult or impossible to identify the words spoken.

Thus, transcription ASR systems typically perform decoding in a complex process integrating phone recognition with the phonetic dictionary to only allow phone sequences corresponding to real words from the dictionary.

Language Modelling

Acoustic recognition on its own is error prone, and even using phonetic modelling expensive to operate.

For large vocabulary recognition *language modelling* has to be incorporated into the recognition process.

Ideal language models would probably model human language processing. However, currently the most effective language models are based on simple statistical models, referred to as n-grams.

A general definition of an n-gram language model is as follows $Pr(w_n|w_1, \ldots, w_{n-1})$, but this model is impractical. Why?

Practical n-grams are bigrams and trigrams, $Pr(w_2|w_1)$ and $Pr(w_3|w_1, w_2)$ respectively.

ASR Transcription Systems

The set of all words which can be recognised by an ASR system is its *vocabulary*.

The total vocabulary available in a typical ASR transcription system includes many words which are very rarely used, but since we never know what is going to be said, they still need to be in the vocabulary!

Thus, we need to have as many words as possible in the vocabulary to ensure that most of the words are present.

ASR Transcription Systems

So called *out-of-vocabulary (OOV)* words are words not present in the vocabulary of a particular ASR system.

OOV words cannot be definition be recognised correctly by the ASR system.

OOV words are usually proper nouns, technical jargon or slang.

This has implications for retrieval. since many useful words for information retrieval are proper nouns or domain specific items of vocabulary.

Since recognition at the phone level also means we do not have to explicitly build models of all words, new words can be added to the vocabulary easily if their phonetic structure is known.

Adding new words to the language model is more difficult - no training data!

Read vs Spontaneous Speech

ASR accuracy varies great depending on the factors outlined previously.

In addition, the recognition accuracy greatly depends on the formality of the structure, level of spontaneity, and location of the speaker relative to the microphone.

Scripted read speech in a quiet environment with a suitable domain specific vocabulary with well trained acoustic and language models is now almost 100% accurate.

Unscripted spontaneous speech in a noisy environment where the speaker is moving around relative to the microphone might be lee than 50% correct.

One of the majpr issues is disfluencies, e.g. false starts, self corrections, “um”, “err”, etc, and poor linguistic structure.

Output of Speech Recognition

Ideally speech recognition systems PLS would produce 100% accurate output. However, as already explained ASR systems make mistakes.

For transcription systems these errors can be classified into the following types:

- substitutions - where the word spoken is misrecognised as a different word
- deletions - where a word spoken is missing from the output, e.g. where 4 words are spoken and the transcribed output only consists of 3 words
- insertions - words not spoken are inserted into the output, e.g. when 3 words are spoken, the transcription system outputs 4 words

Output of Speech Recognition

Deletions and insertions can occur since the speech recognition system has to determine the number of words spoken, and does not always do this correctly.

All types of errors can occur where the acoustic models fail to model the spoken speech input well;

or the language model does not model the spoken word sequence well.

All types of errors can be made worse when an OOV word is spoken.

OOV words cannot by definition be recognised correctly, but their presence can also make it difficult to correctly recognise adjacent spoken words.

Video Mail Retrieval

Video Mail Interface Application.

Video Mail Retrieval

The horizontal bar beside the each message represents the matching score between the message and the search query relative to the highest scoring message at the top of the list.

The user can select any document for playback using the VMR Browser application.

Browsing

Due to the temporal nature of audio, browsing to determine relevance and access relevant information is a significant issue for multimedia data, particularly audio data.

With text we can visually scan a large amount of data very quickly looking for relevant details - search terms can by highlighted in the text to make this even easier.

However, with audio data the scanning process is much slower since only a single stream can be listened to at a time.

Experiments have shown that speech can be played at around double speed (digitally to avoid pitch shifting) without loss of intelligibility, but no faster, but this is cognitively demanding, and the listener will rapidly get tired.

Browsing

Since we are unable to spot the interesting section of a document quickly by looking at it, it must be auditioned from start to finish to find the information required, which takes significant time.

Visual tools for browsing spoken documents are thus potentially very important in maximising data access efficiency.

Potentially interesting portions of the document can be identified and playback can begin at any point by selecting it.

Browsing

Since reading text is faster than listening, one option can be to present users with the ASR transcript of a spoken item.

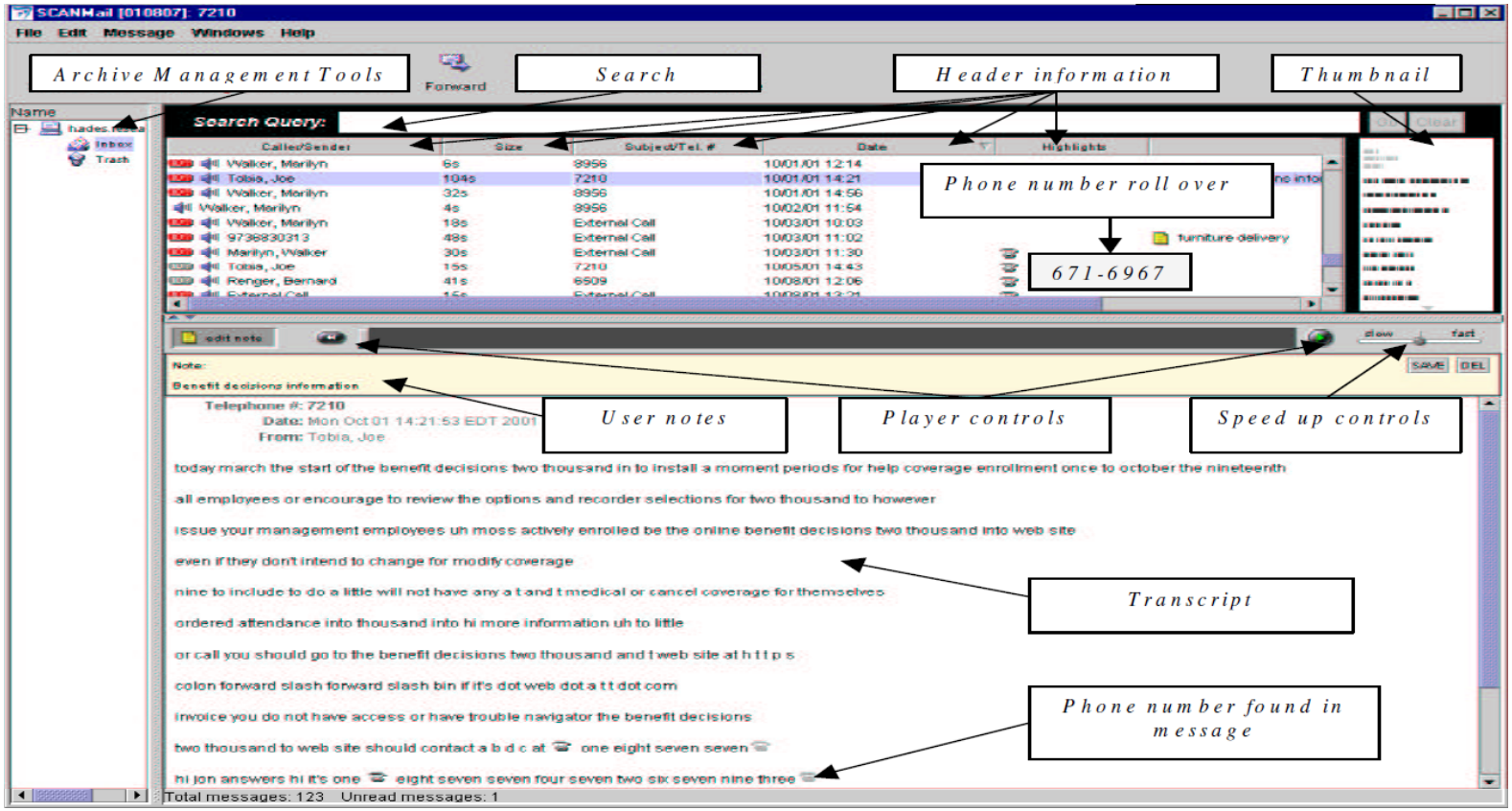
Scanmail is an example of a system that presents the user with the ASR transcript directly, in this case a voicemail message.

A user study suggested that this feature was appreciated. However, these is a relationship between the error levels of ASR transcripts and their usefulness in the system, with higher error rates being less useful.

A user study on the usefulness and usability of ASR transcripts for a web archive found that:

- transcripts with WER $> 45\%$ were unsatisfactory,
- while transcripts with WER $< 25\%$ were useful and usable.

Browsing



ScanMail user interface

Browsing

It is important to keep in mind, that an SCR system must not allow users to develop an unfounded trust in the ASR transcripts.

One study showed that professional users were found to have significant confidence in the the transcripts and their own ability to work with them.

This resulted in the users failing to seek relevant content not explicitly appearing the transcripts, meaning that they missed relevant material if this absence was caused by transcription errors.

The same effect was reported in users of Scanmail.

Recall is more critical for voicemail search and misplaced trust in the ASR transcripts caused users in the study to miss crucial information that was not recognized by the ASR system.

Browsing

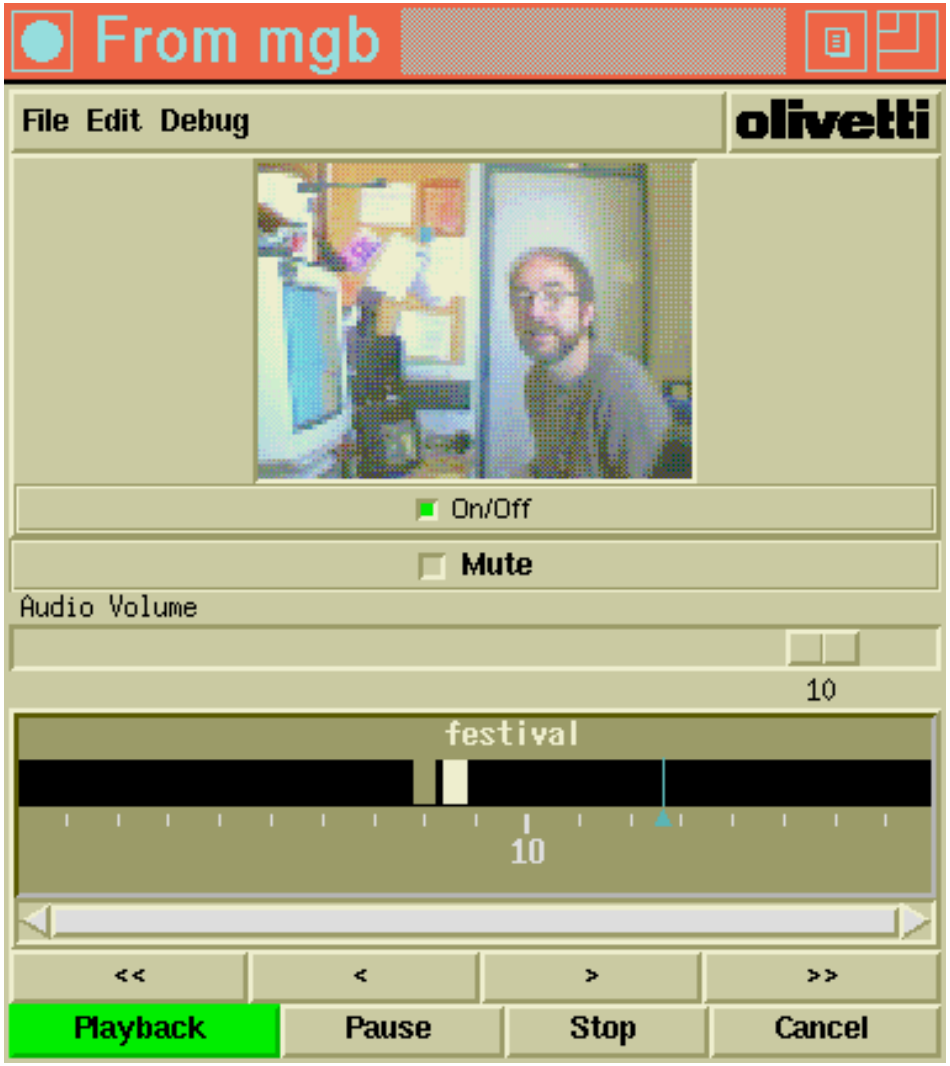
An alternative to use of transcripts is provided by visual tools for browsing spoken documents.

Visual browsers typically attempt to represent the audio stream as a static image that can be viewed at a glance.

Content can be represented on a horizontal left-to-right timeline, with the search word hypotheses displayed graphically along it.

Potentially interesting portions of the document can be identified and playback can begin at any point by selecting it.

Browsing


From mgb
File Edit Debug
olivetti
On/Off
Mute
Audio Volume
10
festival
10
<<
<
>
>>
Playback
Pause
Stop
Cancel

Browsing

The Video Mail Retrieval (VMR) browsers shows words from the query along a timeline.

The brightness indicates the confidence of the ASR system in the correctness of the word.

Browsing

The Dutch podcast search engine *Kunststofzuiger* developed at the University of Amsterdam illustrates typical strategies for SCR players.

The player page has a query-independent representation of the episode, in the form of the podcast title, broadcast date and description, and also a term cloud that has been extracted from the transcript of the podcast.

It also has a query-biased representation of the episode in the form of the player, which contains markers pointing to the moments within the podcast at which the query word occurs.

Clicking one of the markers moves the user to the point in the speech stream at which the query word is spoken.

Browsing

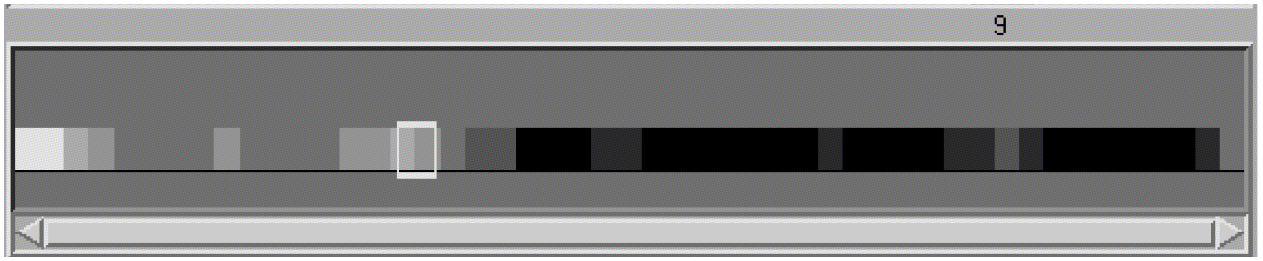

Kunststofzuiger – Podcast details
Back Forward Reload Stop Home http://pir.schuurman.com/kunststofzuiger/index Go Google Search
Post to CiteULike post to del.icio.us del.icio.us MultiMatch Main Web SpeechRetrievalTeam English-Dutch Onlin... IlpsSeminar
iach wiki
Kunststofzuiger – Podcast details
KUNSTstofzuiger
Motes of note from Dutch radio program Kunststof
Zoek!
(i.e. Kunst, Arie of Willem Koning)
Nausicaa Marbe en Jeroen Vullings
Uitgezonden op: 11-03-2008
Aan de vooravond van de Boekenweek praat Kunststof een uur lang met schrijfster Nausicaa Marbe en literair criticus Jeroen Vullings (Vrij Nederland) - die in het dagelijks leven geliefden zijn - over hoe goed of hoe slecht het gaat in de wereld van de literatuur. Plus een reportage met de Amsterdamse boekhandelaar Ton Schimmelpennink. Presentatie Jellie Brouwer
nederland (0/7)
DD: 23
Alternatieven
Sara Kroos 10-03-2008
Edwin Kats 07-03-2008
Mario Molegraaf 06-03-2008
Huub Oosterhuis 05-03-2008
acht actualiteit amsterdam angst antwoord bekend
boek boeken boekhandel cijfer
communicatie dag dezelfde geloof geval helft honderd
jammer jong jongeren kijk kilometer kinderen

Browsing

Two variants on these strategies are:

- A term cloud spread out along the player to give the user a general idea of the topical development over the course of the speech media.
- A heat map display that uses shading or colour to reflect the relative likelihood of a position along the timeline being relevant to a query, rather than showing position of specific words.
  - This approach is adopted in the VMR Broadcast News browser.
  - In order to create this representation, the ASR transcript is divided into equal-length segments each of which is scored against the query.

Browsing



VMR news browser

Segments are scored against the query.

Brightness of segment in display indicates the score against the query.

User can click at any point on timeline to begin playback at this point.

Browsing

The player bar of the Radio Oranje application illustrates the use of a magnifying glass metaphor. (Similar to the fisheye view introduced in the section of the module notes on Hypertext.)

The player displays the entire speech in a timeline, a swell as a magnified view showing a window of 45 seconds around the current position of the cursor.

Above the playbar, the transcript of the currently playing segment is displayed, with the query words in bold and a moving underline tracking the progression of the playback.

The magnified view makes it possible to also depict segmentation information for the entire program in a relatively compact space without losing detail.

Browsing

Radio Oranje using magnifying glass metaphor.

Spoken Input to Applications

The increased computing power available via cloud-based services means that it is now possible to provide online robust automatic speech recognition (ASR) services.

The ubiquity of networked mobile computing devices such as smartphones and tablets which can easily be connected these services, means that ASR services can be accessed on such devices.

Use of speech-based input is often an attractive and efficient option on mobile devices due to the small size of these devices making text input less convenient than on conventional computers.

- However, it must be accurate and fast enough for users to find it more attractive as an option than typed text input.

Spoken Input to Applications

Speech input is not suitable in all situations, for example in public places where use of a speech interface would remove privacy or disturb other people.

An attractive application for voice input on mobile device is spoken input of search queries, referred to as *Search by Voice*.

Search by Voice

ASR for search by voice should be able to support any query that can be entered using text input.

ASR systems for search by voice have potentially huge amounts of training data available:

- Acoustic models: train using manually labelled data (as for a standard ASR system) (referred to as *supervised* data), but also using data from real spoken queries while system is operational (run ASR system on data and automatically extract words with high conference as additional training data tu update models (referred to as *unsupervised* data).
- Language models: use text collections (as for a standard ASR system) and user text queries.

Search by Voice

Use of location information is found to be important to recognition quality, e.g. if the ASR systems knows that the query was spoken in Ireland, then Irish related vocabulary items can be favoured in the ASR output.

Analysis of a queries logs from Yahoo! mobile searched showed:

- Ave length of spoken queries = 4.2 words
- Ave length of text queries = 3.2 words

Spoken queries have more function and questions words (often those words treated as “stop words” in IR systems).

Multi-modal Interaction on Mobile Devices

The large hi-res screens on smartphones and tablets enable use of high quality visual output.

In general, applications can support mulit-modal interaction:

- Input: text, speech, click
- Output: text, speech, graphics, image, video

Multi-modal Interaction on Mobile Devices

Advantages of visual over audio output:

- Visual output rich and efficient presentation - visual bandwidth is much greater than acoustic.
- Lower time to search for and digest information.
- Reduced cognitive load for the user - speech is lost once is it spoken, information must be remembered; visual content can be scanned repeatedly.

Example of effective multi-modal application: Google maps

Visual output not always best: applications where users are “eyes busy” on other activities, users with limited or no visiion.

Content-Based Retrieval of Visual Media

Content-based retrieval of visual media is much more challenging than spoken content retrieval.

Fundamental questions include:

- How should the contents be described for retrieval?
  - For spoken content, we can use words, for visual content, we should use ...?
- What features to describe the content can be extracted automatically?
- What can be labelled manually?
- How should visual queries be posed - text or visual examples?