

# Interactive Media Retrieval in Mobile Communication

Robert van Kommer

“By three methods we may learn wisdom: First by reflection, which is the most difficult; second, by imitation, which is easiest; and third, by experience, which is bitterest.”

“Pour inventer  
on a besoin  
d'expérience.”

“Es ist nicht genug zu wissen,  
man muss auch anwenden können.”

Erstellung als Wissens

# Abstract

<http://diuf.unifr.ch/diva/3emeCycle07/>

- All-in-one mobile phones have changed our social communication behaviors and infotainment habits. For people on the move, accessing media content represents new challenges and different use cases: on the one hand, mobile phones' display and keyboard are much smaller than those on regular PCs; however, on the other hand, these devices are always on, personalized and carried around.
- In this context, the following topics are addressed: how to enhance user's access by cross-media indexing and, furthermore, how could the search/retrieval performance be improved with a "human in the loop" personalization algorithm? Both topics will be illustrated through an interactive media application tailored towards mobile user experience.



## Presentation outline

- Context: horizontal approach in communication
- To enrich multimedia content
- Human-in-the-loop retrieval algorithms
- The multimodal stack widget and it's demo

# Vision: the horizontal approach



# User-centric innovation guidance

- Users stay in the center of innovation and its adoption
- Users need unified, natural service interfaces that are easy to use everyday, everywhere
  - Multimodal and personalized communication
  - Intelligent services to ease the access and to improve service and security



# Horizontal approach in media distribution channels

- The media barriers are disappearing
- Web and IP communication enable a “media agnostic” approach
- What it offers:
  - Unparalleled user experience and interactivity

**CASH**daily

**BILANZ**  
Das Schweizer Wirtschaftsmagazin

**20**  
minuten

Sie wissen das  
Neuste zuerst.



**SF** SCHWEIZER  
FERNSEHEN  
**tsr** télévision  
suisse

**CNN**  
THE WORLD'S NEWS LEADER

You **Tube**

# To Enrich Multimedia Content



“By three methods we may learn wisdom: First by reflection, which is the best; second, by imitation, which is easiest; and third, by experience, which is bitterest.”

“Pour inventer, il faut des connaissances et de l'expérience.”

“Es ist nicht genug zu wissen, man muss auch anwenden können.”

Erstellung als Wissens

# Vision: to enrich content for boosting retrieval capabilities

- **The easy way:** For new content, the horizontal approach suggests to include all valuable “parallel” data at the start



- **The hard way:** For existing content, most operations have to be processed by automatic tools

# Examples of parallel data *with strong/weak synchronizations*

- (strong) Written media and its recordings or Text-to-Speech rendering (Bilanz demo example)
- Audio books (Harry Potter)
- TV Broadcasts and teletxt information
- Radio and speech recognition ([www.audioclipping.de](http://www.audioclipping.de))
- Movies with subtitles (Most DVDs)
- Spoken presentations and their slides
- (weak) Audio, Video, Web content (Background information is used e.g. for language modeling)
- SMIL indexing information

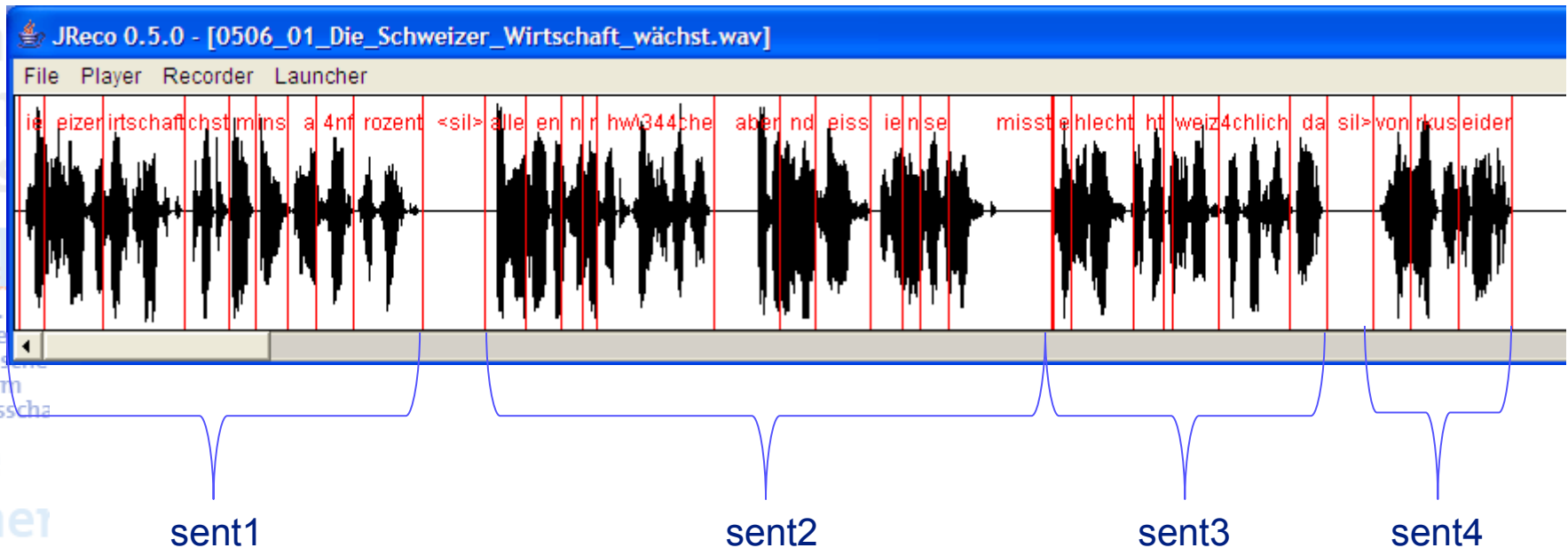
# Text processing steps...

- Natural language processing
  - Text normalization in German
    - UBS → as it is spoken
    - 1. → as it is spoken (erste, ersten,...)
    - 0800 800 800 or 026 400 03 70
  - Language identification (“Guisanplatz”)
  - Text-to-phonetic translation
  - Sentences splitting
  - Text structures identification for automatic SMIL indexing

# Acoustic processing step...

- Creating a new speech recognition model of the speaker (in the general case one would use speaker adaptation)
- Forced-alignment of both text and acoustic representations by using a speech recognizer
  - Depending on the media to be processed: Multimodal processing will be necessary to extract every piece of needed information

# Bilanz demo example



- Sent1: Die Schweizer Wirtschaft wächst um eins Komma fünf Prozent
- Sent2: Alle reden von der Wachstumsschwäche, aber niemand weiss, wie man diese misst.
- Sent3: Wie schlecht steht die Schweiz tatsächlich da?
- Sent4: Von Markus Schneider



## Improved content retrieval

- Rich content is:
  - A text representation and a synchronized audio/video representation
  - Creating cross-media indexing tags
  - Surfing audio/video with text input
  - Surfing text via speech input

# Advanced retrieval and navigation

- Web surfing of multimedia content
  - Retrieval of topics at the sentence level
  - Navigation at the sentence level
    - E.g. Move to the next sentence
    - Retrieve a sentence where ...
- Navigation improvement
  - Introduce audio hyperlinks within video by a localized voice conversion of the original speaker's voice

# Human-in-the-loop retrieval algorithms

“

“By three methods we may learn wisdom: First by reflection, which is the most difficult; second, by imitation, which is easiest; and third, by experience, which is bitter.”

“Pour inventer, il faut des connaissances, de l'expérience et de la persévérance.”

“Es ist nicht genug zu wissen, man muss auch anwenden und lehren.”

vorstellung als wissens

# Vision: Create one meta-user model instead of meta-data models

- The goal is to model the individual user and not data (horizontal approach → to be independent of data)
- Boosting the learning efficiency in order to reduce the number of user's interactions (clicks) and making the process as transparent as possible to the user
- Now, given that meta-user model, we can add intelligence to service interaction we could even make the service proactive.

# Review: human-in-the-loop algorithms

- Post-rating/ranking of traditional keyword search engines
- Inductive learning (SVM <http://svmlight.joachims.org/>)
- Transductive learning (<http://svmlight.joachims.org/>)
- Active learning
- Online learning
- Ranking learning
- Reinforcement learning

The bottom line: How many clicks are needed?

Ultimately, no clicks are needed when the service could proactively anticipate user's needs

# Simulation context and results

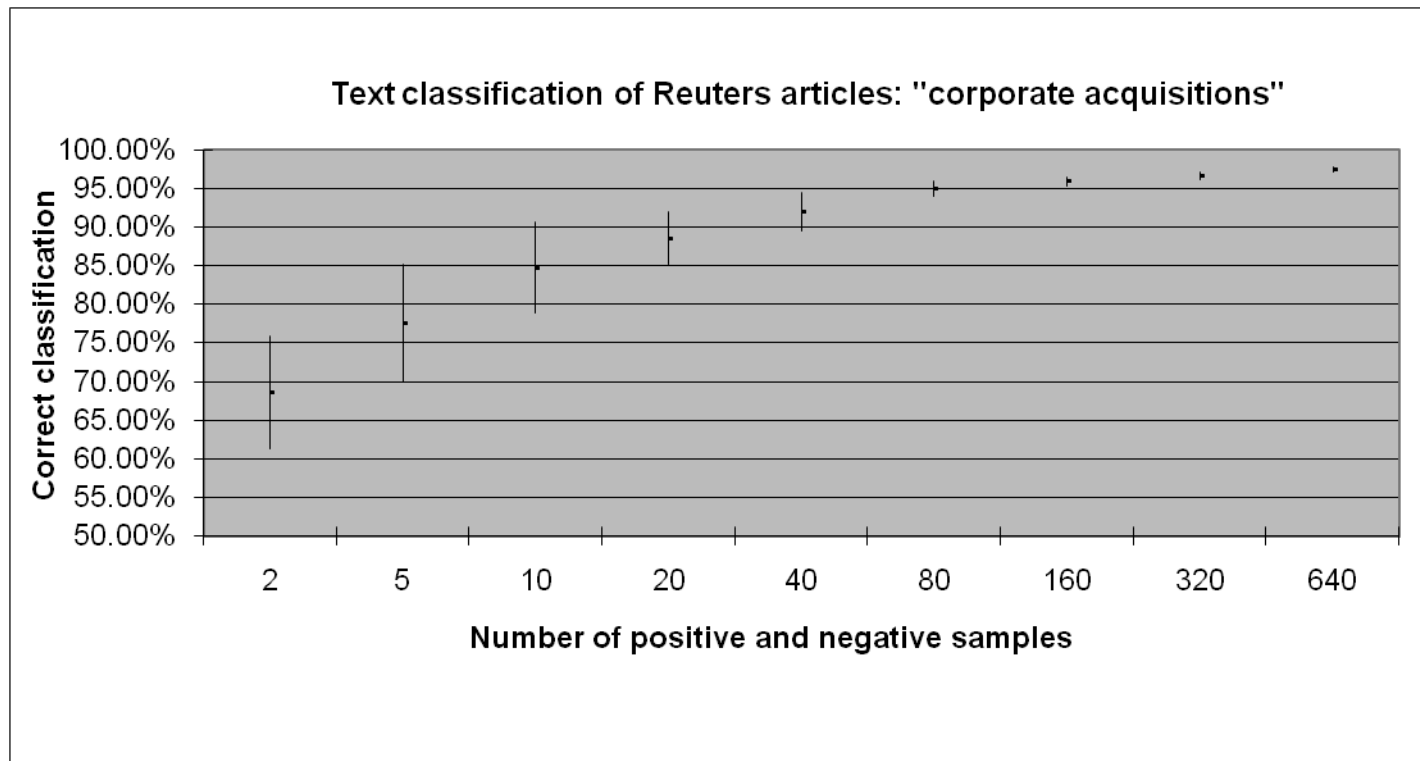
<http://www.daviddlewis.com/resources/testcollections/reuters21578/>

- The task is to learn which Reuters articles are about "corporate acquisitions".
- In the training set, there are 1000 positive and 1000 negative examples.
- The test set contains 600 test samples (300 positive and 300 negative samples).

# Support Vector Machine (SVM)

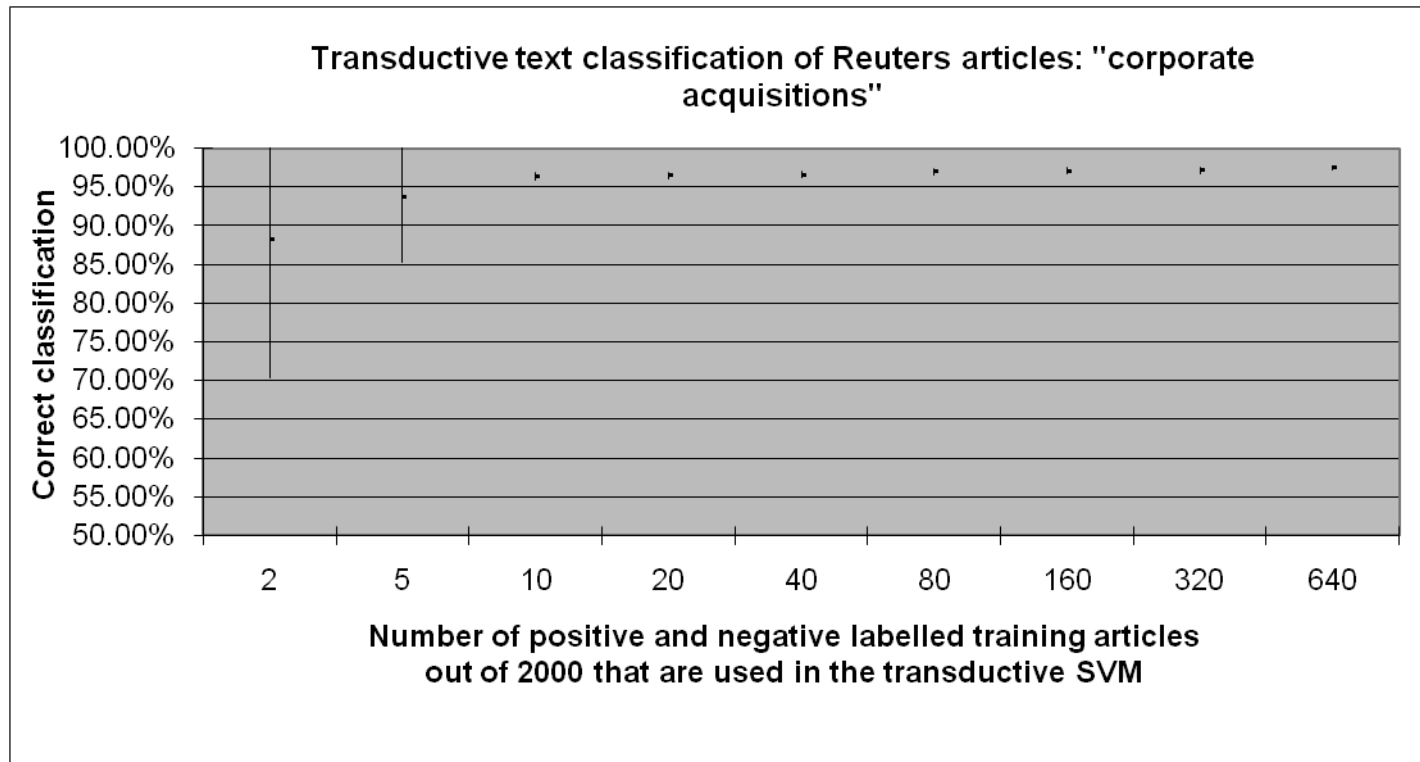
## Inductive learning

### The number of user's inputs needed



# TS VM transductive SVM

## The number of user's inputs needed



# Active learning

- Pool-based (e.g. Tong and Koller)
  - Each learning candidate is selected out of a pool of unlabelled samples; the most critical sample is chosen first, to speed up the training and to reduce human interaction
  - However, data must be available before
- Stream-based (e.g. D. Sculley)
  - On each incoming sample, the algorithm could request human interaction to update the classifier
  - Data is not available before (e.g. incoming e-mails)

# Online learning

- Speed up the learning
  - From the neural network learning paradigm: online learning versus batch-mode learning
  - In our context: The purpose is to learn as fast as possible by using every available sample as soon as possible
- Computation efficiency
  - To reduce the learning time for large training streams

# Improve ranking of results

- To improve the ranking of retrieved results
  - Given a certain number of queries
  - Given a certain number of selections (re-ranking)
  - Given a set of extracted text features
  - The algorithm learns a better ranking
  - See STRIVER <http://svmlight.joachims.org/>

# Demo: Multimodal Stack Widget

**BILANZ**  
Das Schweizer Wirtschaftsmagazin

“

“By three methods we may learn wisdom: First by reflection, second, by imitation, which is easiest, and third by experience, which is bitterest.”

“Pour inventer, il faut résister à l'expérience.”

“Es ist nicht genug zu wissen, man muss auch anwenden können.”

Erstellungskräfte als Wissens

# Vision: a “media agnostic” approach

- Motivation: To make surfing and retrieval of any multimedia content as easy on mobile devices (or easier) than on PCs
- The hard challenge: To cope with the limitations of mobile devices e.g. a small screen and a tiny keyboard.

# Demo overview



Read or/and listen  
Type or/and speak



## Ads

### Reading...

- **Börse Kein Platz für Bären.**
- **Die Strategen der grossen Bankhäuser versprechen allesamt steigende Aktienkurse. Ein Warnsignal? \_\_ So** einzig waren sich die Börsenauguren schon lange nicht mehr: 2007 wird für Aktienanleger ausgesprochen positiv... [...]

### Listening...

Recorded Quality

Text-To-Speech

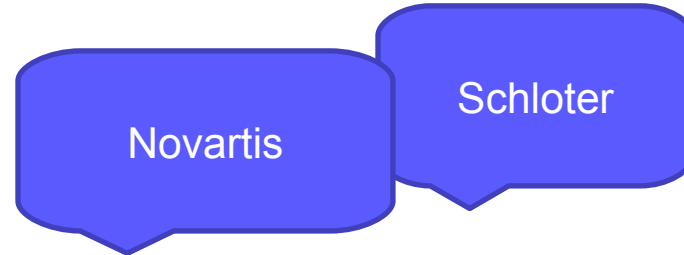


# Bringing all pieces together...

- Integration of parallel data streams
- Integration of intelligence by using human-in-the-loop algorithms
- Integration of a voice search (speech recognition)
- Finally, integration of all interactions into the concept of the **multimodal stack widget**



## Speech recognition integration



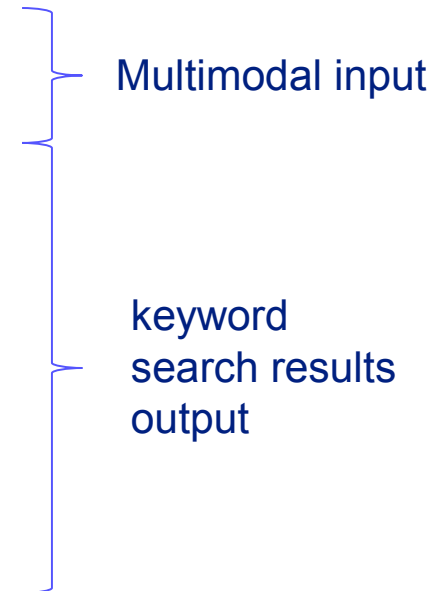
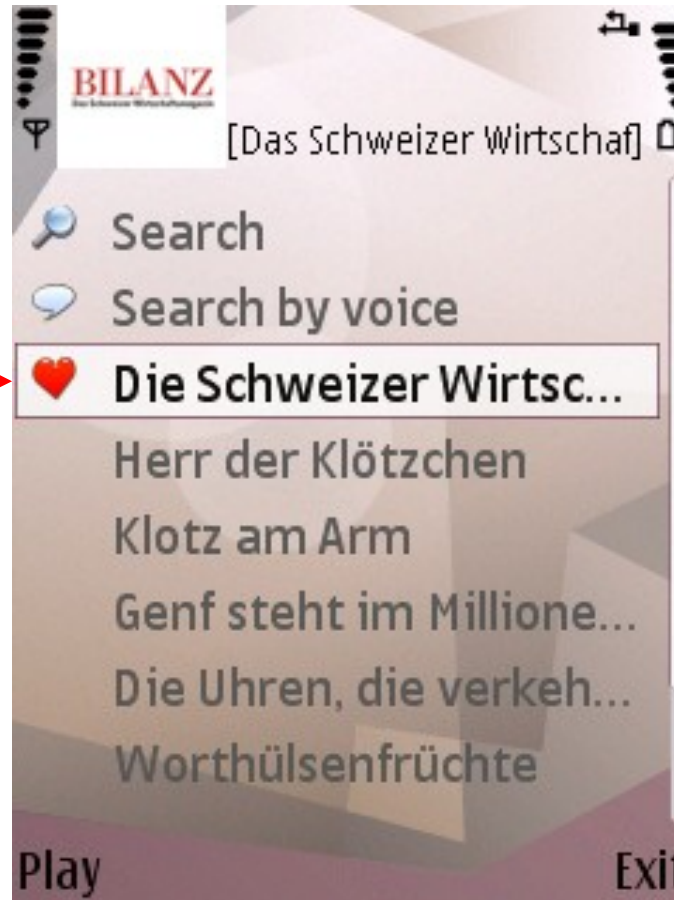
- All-IP server technology from the university of Fribourg
  - With a push-to-talk on mobile phones input
  - Standard open source products
    - Apache Tomcat and Snhinx 4

**DIVA**  
*WebWriteIt!*

<http://diuflx77-vm04.unifr.ch:8080/diva-webwriteit>

# Multimodal stack widget

Human-in-the-loop  
learning result  
(Recommendation)



# Voice search

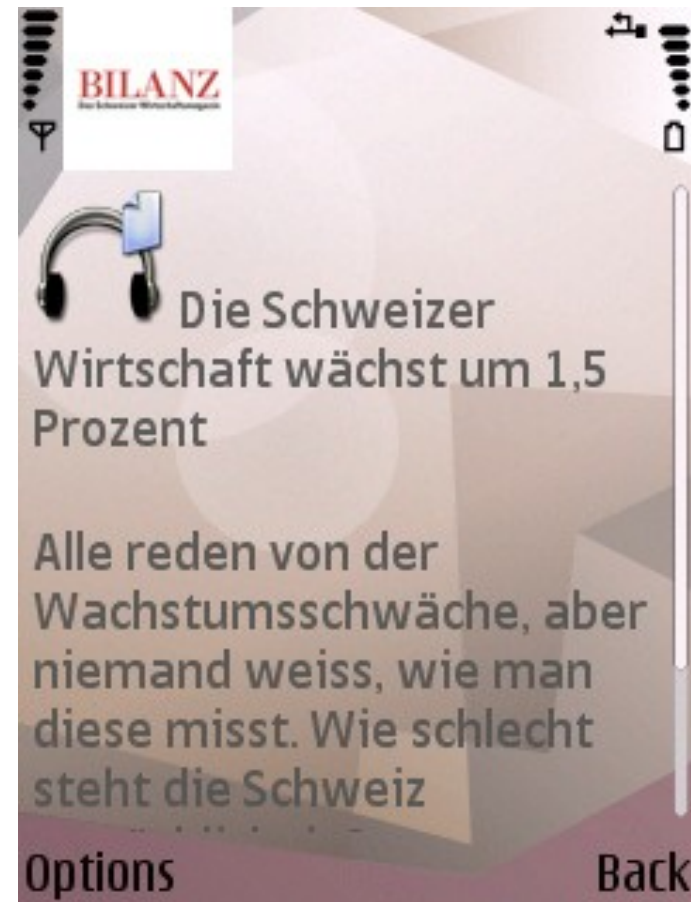
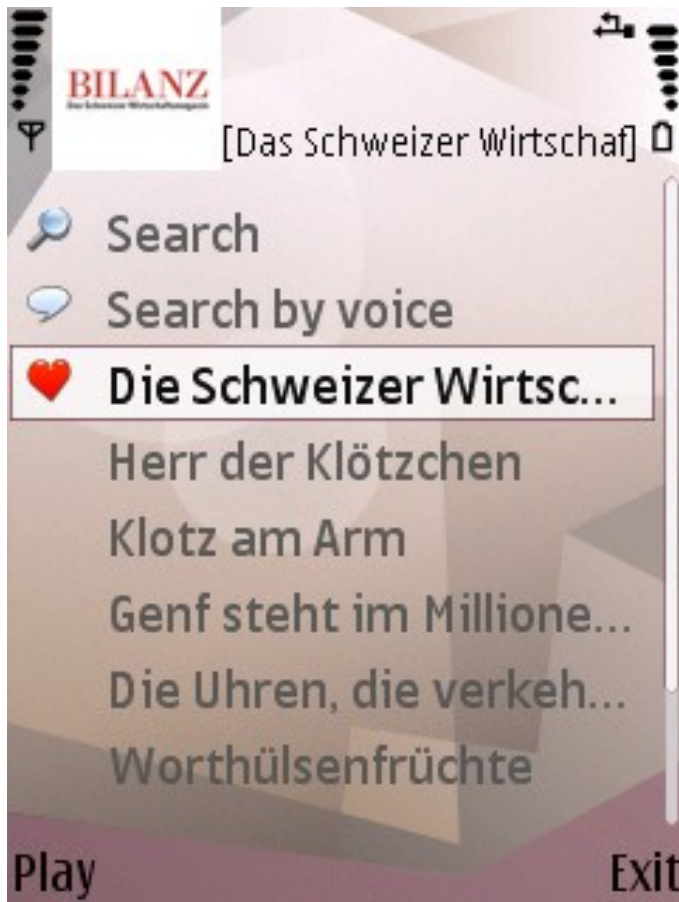


Multimodal search

N-Best speech  
recognition results

keyword search results

# Multimodal stack widget



# Further potential improvement

- Multimodal auto-completion
- Incremental and personalization of the voice search indexing
- Adding audio hyperlinks
- Enabling sentence-based audio navigation

## Starting the UI design

- Applying the user-centric design process
- Running the necessary usability tests

## Conclusion messages

I have presented a **horizontal approach** to enhance multimedia retrieval through an example of an intelligent service



Thank you for listening!

# Voice search at Swisscom

