LREC 2006 workshop: Crossing media for improved information access

The iFinder audio-visual indexing framework for cross media applications

23.05.2006

Dr.-Ing. Joachim Köhler

Head of Research of Competence Center NetMedia

http://www.imk.fraunhofer.de

Joachim.Koehler@imk.fraunhofer.de



Fraunhofer Institut Medienkommunikation



Institut

Medienkommunikation

aunhofer

Outline

- 1. Introduction
- 2. Architecture of iFinderSDK
- 3. Media indexing modules for speech and video
- 4. Generic iFinder system (MPEG-7 metadata extraction and retrieval system)
- 5. Project: AGMA (BMBF)
- 6. Project: Piavida (BMBF)
- 7. Project: WDR/DW Audiomining
- 8. Project: Boemie (IST)
- 9. Project: Live (IST)
- 10. Summary and outlook on future research issues





iFinder - Introduction

Increasing demand on multimedia indexing tools for several applications areas:

- Broadcast domain
- Archiving of parliament speeches
- Media monitoring
- Internet content (podcasts, videocasts)
- Personal recordings (digital camera, Personal Video Recorders)
- etc...

Technical objectives:

- Providing media indexing solutions for practical usage
- Building real life media indexing solutions

Scientific objectives:

- Support of mid- and highlevel features for multimedia search
- Investigation on multimedia semantics
- High indexing accuracy by optimizing the recognition modules to the target domain
- Domain adaptation (creating and usage of huge corpora)
- Investigation of multimodal fusion approaches





iFinder - Introduction

iFinder framework:

- iFinderSDK (toolbox of multimedia indexing modules)
- iFinder (retrieval application)

Flexible multimedia indexing system:

- Support of common meta data formats (i.e. MPEG-7)
- Extendible media indexing capabilities (integration of new indexing methods)
- Simple and open interfaces to allow reuse and integration of the system
- Should be used and enriched in research projects
- Should be used industrial projects and applications (especially iFinderSDK)





iFinder - History

History:

- BMBF research project AGMA (199 2003)
- BMBF research project Piavida (2000 2003)
- Definition of iFinder architecture (2002)
- Programming of iFinder framework (2002 2005)
- Building several applications and mock ups:
 - IBC 2002, 2004
 - · Cebit 2003, 2004
- Prototype for German Parliament
- First industrial project: WDR/DW Audiomining
- IST Project: SHARE (2004 2007)
- IST Project: LIVE (2006 2009)
- IST Project: Boemie (2006 2008)
- IST Project: CitizenMedia (2006 2009)

- ...





iFinderSDK: Basic Architecture

Metadata and media data are separated streams/files

Currently file based processing

Integration with C++ header files

Common processing interfaces for media Indexing Modules

- Init()
- Process()
- Exit()

XML metadata handling:

- MPEG-7 descriptions
- Descriptions are enriched step by step
- Xerces parser is used to guarantee MPEG-7 conformance

Parameters:

- Each module can be configured by XML parameter file



Open Source Technology, proprietary development Signal and pattern recognition processing routines





Media Analysis: iFinderSDK Toolbox





Dr.-Ing. Joachim Köhler

ІМК

Basic recognition system: ISIP recognizer

- LVCSR
- MFCC components
- HMM based (Xword triphones)
- Viterbi decoding
- N-Gram language models
- Lattice output

Training of the Acoustic models:

- German broadcast training database (several hours)
- Semi-automatically labeled and annotated
- X-word triphones are trained
- LM trained with SRI toolkit
- To create training database Transcriber is used

Usage of Recognizer

- LVCSR: 100K word recognizer
- Syllable based recognizer (about 5000 syllables for German language)
- Keyword recognition
- Speech text alignment





Usage of Transcriber: Segmentation of Broadcast Signal

Transcriber 1.4.2		
Edit Signal Segmentation Options Help		
	report	
TD=2 SPEAKER=1 SNS=1 female D=8.5 M=2		
DE3 SPEAKER=0 SNS=0 D=2.4 M=0		
TD=4 SPEAKER=1 SNS=1 female $D=267.0$ M=2		
DID=5 SPEAKER=0 SNS=0 D=16.0 M=0		
DID=6 SPEAKER=2 SNS=1 -male- D=14.1 M=20		
DID=7 SPEAKER=0 SNS=0 D=3.6 M=0		
DID=8 SPEAKER=3 SNS=1 -male- D=26.8 M=5		
ID=9 SPEAKER=4 SNS=1 -male- D=87.3 M=4		
ID=10 SPEAKER=5 SNS=1 -male- D=3.9 M=4		
ID=11 SPEAKER=4 SNS=1 -male- D=13.5 M=4		
ID=12 SPEAKER=5 SNS=1 -male- D=5.6 M=4		
● ID=13 SPEAKER=4 SNS=1 -male- D=30.4 M=4		
DID=14 SPEAKER=5 SNS=1 -male- D=5.7 M=4		
● ID=15 SPEAKER=4 SNS=1 -male- D=17.1 M=4		
● ID=16 SPEAKER=5 SNS=1 -male- D=5.0 M=4		
ID=17 SPEAKER=3 SNS=1 -male- D=19.6 M=5		
ID=18 SPEAKER=6 SNS=1 -male- D=58.7 M=1		
ID=19 SPEAKER=3 SNS=1 -male- D=26.6 M=5		
ID=20 SPEAKER=7 SNS=1 -male- D=76.9 M=1		
DID=21 SPEAKER=0 SNS=0 D=5.5 M=0		
▶ ID=22 SPEAKER=3 SNS=1 -male- D=26.3 M=5		
🕽 ID=23 SPEAKER=8 SNS=1 female D=13.6 M=8		
ID=24 SPEAKER=3 SNS=1 -male- D=27.7 M=5		
ID=25 SPEAKER=9 SNS=1 -male- D=33.8 M=5		
	3303002 ad.trs	
	3303002.wav	
a an	A REAL PROPERTY AND A REAL PROPERTY A REAL PROPERTY AND A REAL PROPERTY AND A REAL PROPERTY	
, α δηματική προσφαιρία του παραγορισμού ματα που ματικά του του του τηματική που του του ματα που του που ματ Ο αναγολογιατική που του του του του του του του του του τ	1986 - 1999 - 1999 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 19 7077 - 1997	······································
a de la desta de la desta de la dela de la desta de		
╡╡╬┟┤╬╪╬┉╫╫┥╊╵╴╸╡╫╊╡╋╡╞╬╴┉╪┉╊╔╛╸╡╶╴╴╴╶╫╴╊╴╧╺╄╴╴┊╺╄╺╫╡┉╴┈╫╡┶╶ ╴	₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩	······································
ID=12 SPEAKER=5 SINS=1 -male-		ID-13 SDEAL/ED-4 SNS-1 male D-3

iFinderSDK: Text-Speech Alignment

Motivation:

- time accurate access to single words
- Generation of training corpus to retrain HMMs

Two Pass Alignment:

- Processing of large audio files (5-10 hours)
- Pass 1: Recognition of important keywords to find coarse alignment
- Pass 2: Grammar generation for each sentence; recognition with grammar; dynamic programming
- Post-Processing: Calculation of posteriorprobabilities to select reliable material

 $p(w | o) = \frac{p'(o | w)p(w)}{p'(o | w)p(w) + p'(o | w_{alt})P(w_{alt})}$

- 20% error reduction with retrained models



Publication:

K. Biatov, J. Köhler: Methods and Tools for Speech Data Acquisition exploiting a Database of German Parliamentary Speeches and Transcripts from the Internet, LREC-2002, Las Palmas, Spain, June 2002



Dr.-Ing. Joachim Köhler



Fraunhofer Institut Medienkommunikation

iFinderSDK: Speech Segmentation Methods

Source material: News show from the broadcaster Deutsche Welle

Wave form file



Segmentation of heterogeneous audio signal:

- Speech/nonSpeech segmentation
- Applause detection
- Jingle detection (HMM based)
- Speaker segmentation based on the Bayesian Information Criterion (BIC)
- Male/Female recognition

Publication:

J. Löffler, K. Biatov, J. Köhler: *Automatic Extraction of MPEG-7 Audio Metadata Using the Media Asset Management System IFinder*, 25th International AES Conference, London, June 2004





iFinderSDK: Speech/nonSpeech Detection Module

Processing Steps:

- 1. Feature extraction of 12 values based on zero crossing parameters
- 2. Classification with multivariate Gaussian classifier
- 3. Optimal smoothing algorithm based on homogeneity calculation (entropy)
 - Find optimal boundaries using dynamic programming
 - Constraints: MinDuration (0,25 second) and MaxDuration(5,0 second)
 - As score value an entropy measurement is calculated:

$$H(s,e) = \sum_{i=s}^{e} m_i \left(\frac{m_i}{N}\right) \log\left(\frac{m_i}{N}\right) \qquad \mathsf{L} = (\mathsf{N}, \mathsf{N}, \mathsf{S}, \mathsf{S}, \mathsf{S}, \mathsf{N}, \mathsf{N}, \mathsf{N}, \mathsf{N}, \mathsf{S}) \\ \mathsf{H} = \mathsf{0} \qquad \mathsf{H} = -0,9$$

Recognition results for speech frames/segments of German parliament data:

	correct recognized	correct recognized
	frames using 2	segments using 3
speech	76.6%	96.7%

Publication:

K. Biatov, J. Köhler: *An Audio Stream Classification and Optimal Segmentation for Multimedia Applications*, ACM Multimedia, San Francisco, November 2003





iFinderSDK: MPEG-7 Meta Data Generation Process

Audio/Video Data



iFinder System - Distributed Media Indexing and Retrieval System



Les Pins, France, December 2002





iFinder: Retrieval User Interfaces (MPEG-7 based Text Video Browser

强 lFinder Client Ver: 2.0

File Options Help

Query Window Results Window Browser Window

Text Video Browser - Streaming



X

etwas anderes Es geht darum dass wir unsere Verpflichtungen auch in anderer Hinsicht ernst nehmen. Woran ist das Bündnis mit den Entwicklungsländern in Bali gescheitert Ich will es Ihnen sagen Es ist daran gescheitert dass sich Europa nicht darauf verständigen konnte der Aussage dass auch wettbewerbsverzerrende und nicht nur umweltschädliche Subventionen abgebaut werden sollen zuzustimmen. Ich will Ihnen an einem einfachen Beispiel erläutern was das heisst, wir die Vertreter der Bundesrepublik Deutschland hatten in dieser Frage eine sehr eindeutige Position. Wir haben ganz deutlich gesagt Die Subeventionen müssen gesenkt werden. Herr Kollege wir sind nicht bereit eine Politik zu akzeptieren die beispielsweise zehn Jahre nach dem Sturz des Apartheidregimes in Südafrika dazu geführt hat dass einerseits südafrikanische Produkte auf dem europäischen Markt angeboten werden durften. Europa ist unter anderem mit Pfirsichen auf den südafrikanischen Markt gegangen deren Vertrieb so stark subventioniert war dass beispielsweise griechische Pfirsiche trotz der Transportkosten in Südafrika zehn Prozent billiger waren als die heimischen Pfirsiche. So hat man eine funktionierende südafrikanische Pfirsichkonservenproduktion kaputtgemacht drei tausend Menschen sind arbeitslos geworden. Von diesen Erfahrungen sprechen die Entwicklungsländer. Gerade im Hinblick auf unsere Kolleginnen und Kollegen in der Europäischen Union sage Gerade im Hinblick auf unsere Kolleginnen und Kollegen in der Europäischen Union sage ich wir werden darüber in der übernächsten Woche im <mark>Umwelt</mark>rat sehr ernsthaft diskutieren müssen Wenn





tior

Research Project AGMA (Automatic Generation of Metadata based on MPEG-7)

Project facts

- BMBF project (1999 2003)
- Support and exploitation of MPEG-7
- Speech and video indexing methods

Research Goals:

- Multimodal indexing (speech and video features)
- Automatic exploitation of very large training databases

Data and domain:

- Recordings of the German Parliament (> 100 hours)
- MPEG-2 streams from German broadcaster Phoenix
- Closed speaker set (>500 speakers)
- Transcripts are available as stenography
- Face database was collected (manually selected and annotated)





Research Project AGMA (Automatic Generation of Metadata based on MPEG-7)

iFinderSDK modules:

- Speech recognition (speech text alignment)
- Speaker segmentation
- Speech/Nonspeech detection (applause detection)
- Cut detection
- Face detection and recognition (50 faces)
- Person segmentation

Retrieval application

- Xindice database
- Client server application based on Corba
- Java frontend
- Web services to access content over Web technogies (browser, VoiceXML)





Research Project Piavida (Personalized Interactive Audio, Voice and Video Information Portals and Applications)

Project facts

- BMBF project (2000 2003)
- Speech and video indexing methods
- Mobile data access

Research Goals:

- Dataminig on multimedia data (spoken document retrieval)
- Multimodal indexing (speech and video features)

Data and domain:

- Recordings from the broadcaster DW (80 hours of "Kalenderblatt")
- Realstreams are recorded and decoded to PCM data
- Transcripts are available and were used to build training database





Research Project Piavida (Personalized Interactive Audio, Voice and Video Information Portals and Applications)

iFinderSDK modules:

- Speech recognition
 - speech text alignment
 - Syllable based recognition
- Speaker segmentation

Retrieval application:

 Web based demo application

🚰 Search Results of Syllsearch - Microsoft Internet Explorer 📃 🔲									_ 🗆 🗙		
] [<u>D</u> atei	<u>B</u> earb	beiten	<u>A</u> nsicht	<u>F</u> avoriten	E <u>x</u> tras	2				A
	Sy We	llabl eb Ir	le Se nterf	arch ace			Fra	aunhofer	IMK Institut Medienkon	nmunikatio	n
	Search Results for Nelson Mandela [n@l zo:n man de: la:]										
	Se	ore	Syl	lables	Date	T	itle	Word	Seg.	Doc.	
	87%		[n@l i man d	ts@n le: 1a:]	11.02.00	1990: Mand	Nelson ela frei	word	<		
e								🌚 Inte	ernet		





Project facts

- Industrial project
- Development of an pilot application for audiomining
- 2005/06

Research and Development Goals:

- Automatic generation of a structure of broadcast recordings
- Indexing and search for word queries (e.g. "Angela Merkel")
- Standalone system

Data and domain:

- WDR and DW provided 160 hours of broadcast recordings (4 different types of radio shows)
- Several hours are transcribed and labelled
- Mixture of speaking styles:
 - Professional and unprofessional speakers
 - Studio and telephone speech
 - Over lapped by music (background music)
 - Dialogue based speaking style versus read speech





Project Audioming (with broadcaster Deutsche Welle and WDR)

iFinderSDK modules:

- Speech recognition, syllable based recognition (using the Piavida models)
- Speaker segmentation and clustering

Retrieval application

- Content based access to audio segments based on syllable based ASR
- Graphical presentation of automatically generated metadata
- Web based system with Flash Audio browser and full workflow integration
- JAVA middleware (i.e. Java servlets)
- Installation on a standard PC





🌮 Getting Started 🔂 Latest	Headlines	
WDR		Fraunhofer Institut Medienkommunikation
D	3303*	
Titel		
Erstsendedatum	von bis	
Rundfunkanstalt	Alle Programmkennzeichen Alle	
Suchworte	Johannes Rau	
Suchschärfe	>= 🗙 80 💌	
Suchart	🔿 Datenbanksuche 💿 Silbensuche	
	Suchen Neue Suche Hilfe	

WDR	Fraunhofer Institut Medienkommunikation
Suchworte: 'Johannes Rau'	
Es wurden 54 Treffer gefunden.	
1. johannes rau, 100%	
D: 3303002	ESD: 02.03.2004
SRTI: Funkjournal Komplettmitschnitt	Rundfunkanstalt: DW
BETI:	Programm: Deutsches Programm
2. johannes rau, 100%	
D: 3303003	ESD: 03.03.2004
SRTI: Funkjournal Komplettmitschnitt	Rundfunkanstalt: DW
BETI:	Programm: Deutsches Programm
3. johannes rau, 100%	
D: 3303011	ESD: 11.03.2004
SRTI: Funkjournal Komplettmitschnitt	Rundfunkanstalt: DW
SETT: 110204 MS 17 05	Programm: Deutsches

Ę

ID: 3303002 BETI: SRTI: Funkjournal Komplettmitschnitt SHTI: 020304 MS 15 05 Rundfunkanstalt: DW, Programmkennung: Deutsches Programm, Erstsendedatum: 02.03.2004 Suchworte: 'Johannes Rau'

					l Tref	fer
Sta	rt Dau	ler Segn	entart Spr	echer		
00:1	0:26 00:	01:13 spee	ch Speaker	_M_ID-7		-
00:1	L:39 00:	00:03 spee	ch Speaker	M_ID-3		
00:1	L:43 00:	00:05 nons	peech			
00:1	1:49 00:	00:25 spee	ch Speaker	M_ID-3		
00:	ll:53 jo	hannes rau	100%			
00:1	2:14 00:	00:14 spee	ch Speaker	_F_ID-8		
00:1	2:28 00:	00:27 spee	ch Speaker	M_ID-3		
00:1	2:56 00:	00:33 spee	ch Speaker	M_ID-9		
00:1	3:30 00:	00:05 spee	ch Speaker	M_ID-3		
00:1	3:35 00:	00:47 spee	ch Speaker	M_ID-9		
> 00:1	4:23 00:	00:11 spee	ch Speaker	M_ID-3		
00:1	4:35 00:	00:54 spee	ch Speaker	M_ID-9		
< 00:1	5:29 00:	00:14 spee	ch Speaker	M_ID-3		
00:1	5:43 00:	00:24 spee	ch Speaker	M_ID-9		
00:1	5:08 00:	00:07 spee	ch Speaker	M_ID-3		
00:1	5:16 00:	00:17 spee	ch Speaker	M_ID-9		
00:1	5:34 00:	00:34 spee	ch Speaker	M_ID-3		
00:1	7:08 00:	00:10 spee	ch Speaker	_F_ID-8		
f[sec] 00:1	7:19 00:	00:16 spee	ch Speaker	M_ID-10		
00:1	7:36 00:	01:04 spee	ch Speaker	_F_ID-8		
00:1:	3:41 00:	00:14 spee	ch Speaker	M_ID-10		
arke 00:1:	8:55 00:	00:07 spee	ch Speaker	_F_ID-8		

~

EU IST Project LIVE (SO: 2.4.7)

Project facts

- IST project (2006 2009)
- Overall project goal: new iTV production methods for live sport content
- IP project (11 Mio. Budget)
- Coordinator: Fraunhofer IMK

Research Goals on metadata extraction:

- Indexing methods for live content
- Semi-automatic annotation tool for live content
- Indexing of ORF sport archive content



EU IST Project BOEMIE (SO: 2.4.7)

Project facts

- IST project (2006 2008)
- Overall project goal: research on multimedia ontologies and semantics
- STREP project
- Coordinator: NCSR

Research Goals on metadata extraction:

- New audio and video indexing methods
- Semantic models for audio visual data
- sport domain





Page 26

Fraunhofer Institut Medienkommunikation

- iFinderSDK provides a set of modules for media indexing
- iFinder is used in several research and development projects
- Research issues:
 - Indexing of live content
 - · Semantic multimedia indexing
 - High accuracy indexing for each module
 - Semi-automatic generation of annotated databases
- Other projects:
 - EU IST SHARE
 - EU IST CitizenMedia

