Speech Processing and Retrieval in a Personal Memory Aid System for the Elderly

A.Sorin[1], H.Aronowitz[1], J.Mamou[1], O.Toledo-Ronen[1], R.Hoory[1], M.Kuritzky[1], Y.Erez[1], B.Ramabhadran[2], A.Sethy[2]

[1] IBM Haifa Research Lab, Israel
[2] IBM T. Watson Research Center, US

Outline

Introduction

- **HERMES EU project**
- **The role and challenges of speech processing**
- **Data collection**

Speech-to-text transcription

Speaker tracking

Spoken information retrieval

Conclusions

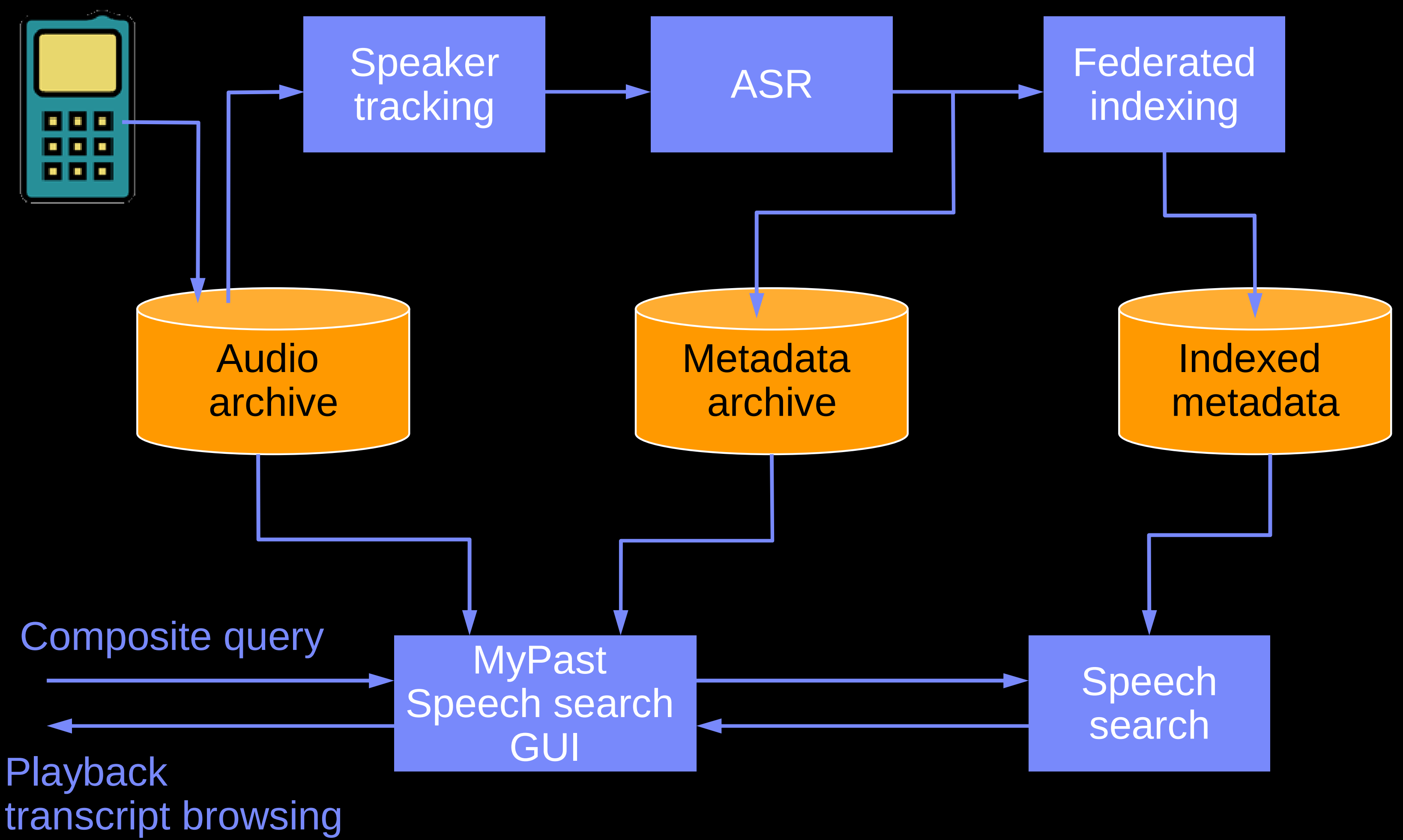
HERMES EU project at a glance

- 3 years long multidisciplinary collaborative project partially funded by EU under FP7
  – 6 academic and industrial partners
- Development of personal assistive system alleviating aging-related cognitive decline through providing *“external memory”* and cognitive training & stimulation
  – Audio-visual data capturing by mobile device (PDA) and stationary video cams at home
  – Metadata extraction
  – Memory support (on-demand, contextual), cognitive games
- HERMES services
  – MyPast – exploring past experience recorded, processed and stored
  – MyCalendar – contextual reminders at right time and place
  – MyTraining – memory exercises based on personal audio-visual data

HERMES MyPast application

- HERMES system displays available audio-visual info relevant to user's query
  - What did the **doctor** tell me **yesterday** about the **diet**?
  - Show conversations with **Paola** about classic movies that we had in **August**

Speech processing & retrieval control flow in HERMES



Challenges to speech processing

- Conversational speech, open domain
- Distantly placed PDA microphone
- Elderly voices atypical to the main stream applications
- Massive data collection for training is unaffordable

Training data collection

- Target language for the system prototype is Castilian Spanish
- Audio data was recorded during a user study at the beginning of the project
  - 47 elderly and 4 young (gerontologists) speakers
  - Simultaneous recording by the PDA and a high quality headset microphones for research purposes
  - 47 dialogues/interviews – 18 hours
  - 182 free style monologues – 9 hours
  - 20 readouts – 13 hours
- All the data passed manual verbatim transcription and speaker labeling

Outline

- Introduction
  - HERMES EU project
  - The role and challenges of speech processing
  - Data collection
- **Speech-to-text transcription**
- Speaker tracking
- Spoken information retrieval
- Conclusions

Speech transcription work in HERMES

- Attila speech transcription toolkit developed by IBM Research
- Two-pass decoding
  - 1st pass – speaker independent
  - Speaker adaptive (VTLN, fMLLR) and discriminative (fMMI) transformations of feature vectors
  - 2nd pass with discriminative (MMI) Acoustic Models
- 3-gram statistical Language Models

- Three development phases
  - Baseline system ASR0
  - Intermediate system ASR1
  - Advanced system ASR2

Baseline system ASR0

- Spanish ASR system developed by IBM in EU TC-STAR project
  – Trained on hundreds of hours of manually transcribed parliamentary speeches
  – 4,000 HMM states, 100,000 Gaussians
  – 8% Word Error Rate (WER) in TC-STAR evaluation
- High mismatch between ASR0 training conditions and HERMES target conditions

Acoustic mismatch (channel & speaker)

Linguistic mismatch

	Lip microphone	PDA
Readout	WER=24%	WER=41%
Dialogues	WER=48%	WER=68%

Intermediate system ASR1

- Language Model adaptation
  – New LM built on 100,000 words subset of HERMES conversations
  – Interpolation between ASR0 LM and the new LM

- Per-speaker acoustic model adaption – speaker enrollment
  – Supervised MLLR-based adaptation of the ASR0 AM on HERMES monologues
  – Yields some compensation of the channel mismatch

Acoustic mismatch compensation

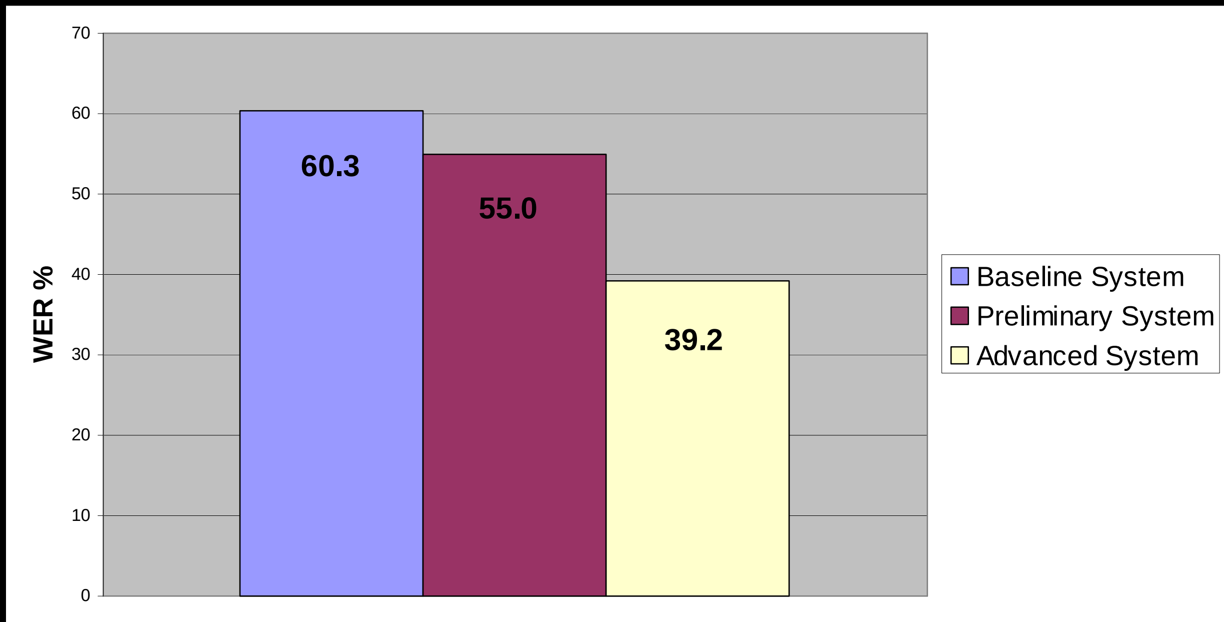
Linguistic mismatch compensation

	Baseline AM	ASR1 AM
Baseline LM	68% (ASR0)	64%
ASR1 LM	60%	54% (ASR1)

Advanced system ASR2

- Trained completely on HERMES PDA data with bootstrap by ASR0
  - 38 hours, 320K words, 49 speakers. ASR0/1: 100 hours, 70M words, hundreds of speakers
  - 1K HMM states, 30K Gaussians. ASR0/1: 4K states, 100K Gaussians
  - Does not require speaker enrollment

Accuracy evaluation on PDA-recorded dialogues and monologues from two speakers



Outline

Introduction

- HERMES EU project
- The role and challenges of speech processing
- Data collection

Speech-to-text transcription

Speaker tracking

Spoken information retrieval

Conclusions

Speaker tracking on two-parties conversation

- Speaker tracking - who spoke when
  - Speaker diarization - segmentation to speaker turns and clustering
  - Speaker recognition - assigning speaker identity to the clusters

- Speaker tracking in HERMES
  - Conversation of the system owner with another person
  - Used for search: *find my conversations with Maria*
  - Enhanced speech transcript readability - who-said-what - for the conversation browsing

Two-speaker diarization

- *H. Aronowitz, "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Odyssey, 2010.*
  - Assumption: speaker characteristics change faster than speaker identity
  - GMM super-vector parameterization of 1 second long super-frames
  - Unsupervised NAP compensation of intra-speaker variability
  - HMM-based Viterbi segmentation
  - 2.8% Frame Error Rate on NIST 2005 telephony data
- Evaluation results
  - 24% Frame Error Rate on the HERMES dialogues

Speaker recognition on dialogues

- Speaker recognition is applied to the imperfect clusters provided by the diarization
- State-of-the-art speaker recognition algorithms suffer from the interfering speaker
  – Features warping
  – Inter-session intra-speaker (IS-IS) variability modeling
  – Score normalization
- Novel approach reduces the influence of the interfering speaker
  – *H. Aronowitz, V. Aronowitz, "Efficient score normalization for speaker recognition", in ICASSP, 2010.*
  – *Y. Solewicz, H. Aronowitz, "Two-Wire Nuisance Attribute Projection", in Proc. Interspeech, 2009.*
  – Equal Error Rate ≈ 4% on NIST telephony data
- 11.3% Equal Error Rate on the HERMES dialogues

Outline

Introduction

– HERMES EU project
– The role and challenges of speech processing
– Data collection

Speech-to-text transcription

Speaker tracking

Spoken information retrieval

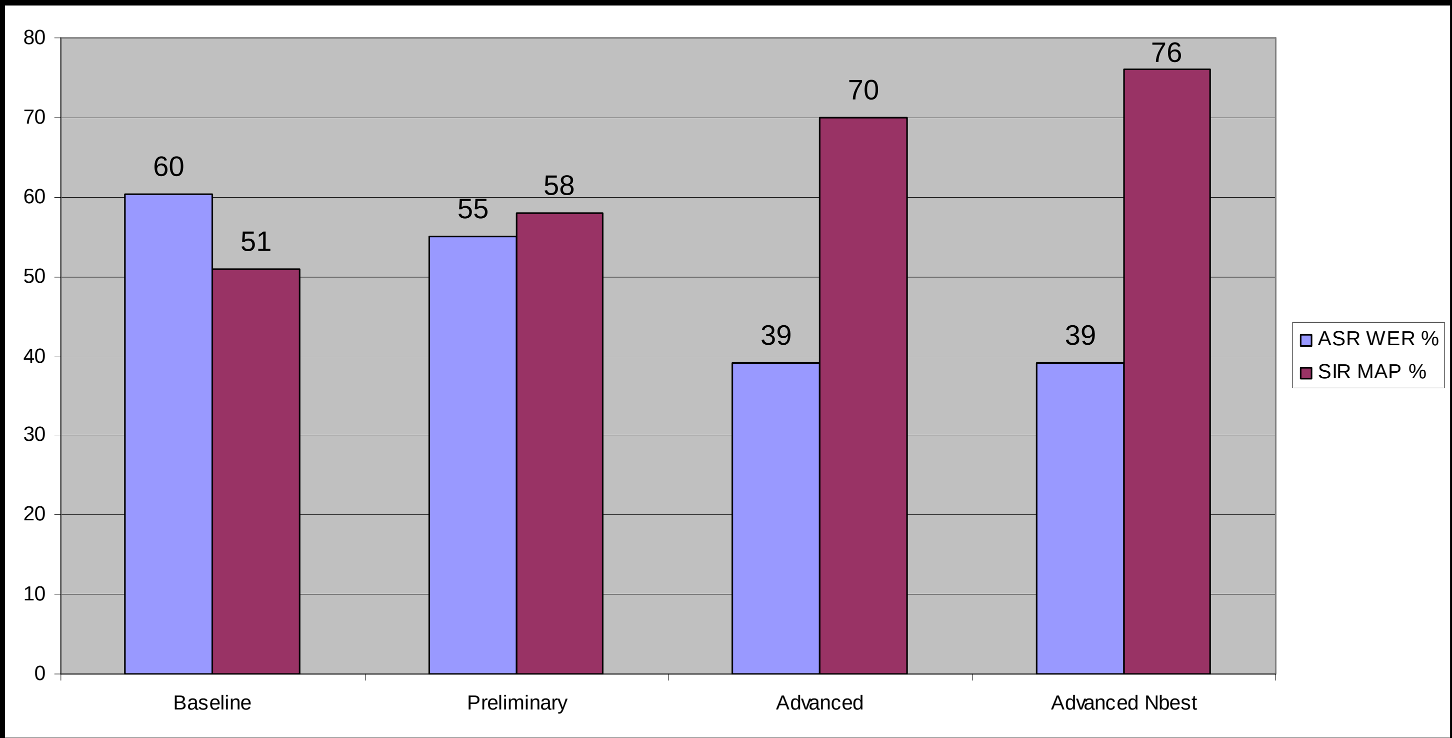
Conclusions

Indexing, query language and search

- The contents of the index repository
  - ASR Word Confusion Networks: N-best word alternatives in stemmed form with their confidence measures
  - Word time stamps
  - Speaker identities
- Query language
  CONTENT:/hermes solution/ SPEAKER:/alex/
  CONTENT:/"hermes solution" AND speech/ SPEAKER:/alex OR hagai/
- Search
  - Returns N top-relevant items (conversation ID, time stamps of the fragment)
  - Spell checking, suggesting a *better* query – "did you mean?"

Spoken information retrieval evaluation

- Spoken conversation retrieval task
- Content-based queries, speaker identity was not used
- 20 conversations used for the ASR evaluation
- 55 manually composed queries
- Ground truth: relevant conversations for each query are found using textual search over the manual verbatim transcripts
- Mean Average Precision (MAP) measure



Conclusions

- Speech processing technologies become mature enough to meet the challenges posed by AAL applications

- Availability of domain-specific data for training is crucial
  - Small system ASR2 trained on relevant data outperforms the big adapted ASR1 system originally trained on irrelevant data
  - Deficit of domain-specific annotated data is typical for a multidisciplinary AAL project
  - Broad collaboration and data sharing is needed, e.g. in the framework of EU FP

- Recent advances in speaker recognition yield reasonable performance on two-speakers conversations recorded by a distant mobile device

- Advanced speech search technology allows to approach the performance of the textual information retrieval despite of substantial ASR errors