Alexander Sorin

26 May 2011



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

A.Sorin¹, H.Aronowitz¹, J.Mamou¹, O.Toledo-Ronen¹, R.Hoory¹, M.Kuritzky¹, Y.Erez¹, B.Ramabhadran², A.Sethy²

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



Introduction

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

Speech-to-text transcription

Speaker tracking

Spoken information retrieval



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**

Back to Home	Screen	📖 MyPast	Start Recording Recording	Inactive			
Friday	Order by: Name Last Visit)	Back to Home Screen	🛵 MyPast	Start Recording Recording	Inactive 🥥 🛄 MyPast St	art Recording Recording Inactive
h o m e		hr00708	Friday 24.09.10 22:00 00:00 02:00 04:00 h	0 06:00 08:00 10:00 1	22:00 Sunday 17.1		20:00 22:00
m			o m e	A B C D E F	h Previous Recording 0 17. Oct 2010, 21:47 e	2	
b i l e			m o b i	G H I J K L M N O P Q R S T U V W X	m o	a the state bady to be in the state	94 10.139
Ang 2010	Clear selection	V Done	•	Y Z Space	b i I		
	People)) Speech	Ang 2020	Clear	e	que ir es muy a de cultura y y manda no alemania	
				People M Speech	◀ 5mp 2018 ◀ 15.1	17. Oct 2010, 22:11:09	. 10 Nov 2010
						People	



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



7 Transcriber 1.5.1								
Elle Edit Signal Segmentation Options Help								
mi hermano alguna vez también ha tenido y no dije nada a nadie		1						
pero un dolor fortísimo a la altura del corazón								
fortisimo	fortísimo							
<pre> <unintelligible></unintelligible></pre>								
speaker #1								
o <ub> te dijo algún médico que era <ub></ub></ub>								
speaker#2								
estaba en una excursión								
y el guía de la excursión me dijo esos son nervios y aires								
speaker#1								
maxies 82								
weare								
 suminaligibles 								
restant!								
e de todos modos te voy a pedir una rediografía te voy a pedir una placa de los pulmones								
a construction of a point and roundyfund on to a point and plane as the planeton								
and a state of the								
w te you a pedir un ecocardiograma que es una como una ecografía de corazón pero hueno no n	orque niense que	tenge						
nada malo de <ub> ni de corazón ni de pulmones <ub> simplemente bueno</ub></ub>	ordeo broupo deo	oongub						
pues por descartar y porque nunca nos viene mal hacernos una revisión de vez en cuando <uh></uh>	no porque piens	e ni que						
te vaya a dar un infarto ni que <uh></uh>		-						
H(+(11 ▶ → →)) 1 hr06408_i02_s2doctorvisit_pda								
		<u>.</u> .						
a faith and an an an an an a state of the st	ويقتله ويستعله							
	International States	A Manual						
l ber ber a the set								
speaker#2	speaker#1	speaker#2						
he notado a hace muchos años y de vez mi hermano alguna vez. pero un dolor fortisi en cuando la la unbuia di en anda a padia del corazón mo cuando la de corazón (en cuando la de corazón en cuando en cuando la de corazón en cuando la de corazón en cuando la de corazón en cuando en cuando la de corazón en cuando e	 <ub> te dijo algún. estaba</ub> <ub> ón</ub> 	. y el guía de l						
3:10 3:15 3:20 3:25 3:30	3:35	unou						





Introduction

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

Speech-to-text transcription

Speaker tracking

Spoken information retrieval



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
 - -2^{nd} 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

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

Acoustic mismatch (channel & speaker)

-inguistic mismatch



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

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

-inguistic mismatch

compensatio



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



© 2009 IBM Corporation



Introduction

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

Speech-to-text transcription

Speaker tracking

Spoken information retrieval



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 \approx 4% on NIST telephony data
- 11.3% Equal Error Rate on the HERMES dialogues



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





- 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 twospeakers 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