INTERACTIVE
MULTIMODAL
INFORMATION
MANAGEMENT

IM2: Audio Processing
Oct. 2008 – Sept. 2009

Key Achievements

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

Overview of IM2.AP
Partners

John Dines (IP head), Phil Garner (Deputy)

Petr Motlicek, Mathew Magimai Doss

Students: Deepu Vijayasenan, Joel Pinto

Gerald Friedland, Nelson Morgan

Students: Mary Knox, Oriol Vinyals

Completed PhD: Kofi Boakye, David Gelbert

Overview of IM2.AP
Projects

- Automatic Speech Recognition
  - Meeting room speech recognition
  - Spoken term detection
- Speaker recognition, segmentation and clustering
  - (Overlapping) speech detection & segmentation
  - Diarization (who spoken when)
- Microphone array processing for signal enhancement and ASR
  - Microphone array beamforming
- High level modelling of spoken language
  - Summarization

Overview of IM2.AP
Goals / milestones

- Long-term goals
  - Conduct fundamental research
  - Application of this research to IM2 domain
  - Dissemination via publication and cooperation with IM2 partners
- This year: strong focus on working prototypes
  - Requirement for real-time, online processing
  - Both intra- and inter- IP integration effort
  - Dissemination of software tools

Key Achievements
Automatic Speech Recognition

- Participation in NIST Rich Transcription evaluations
  - Continued improvement on the multiple distant microphone task (MDM)
  - Paradoxically, individual headset microphone task (IHM) keeps on getting more difficult

ICSI-SRI topped IHM task, Idiap-AMIDA MDM task

  - Reversal of previous years' results

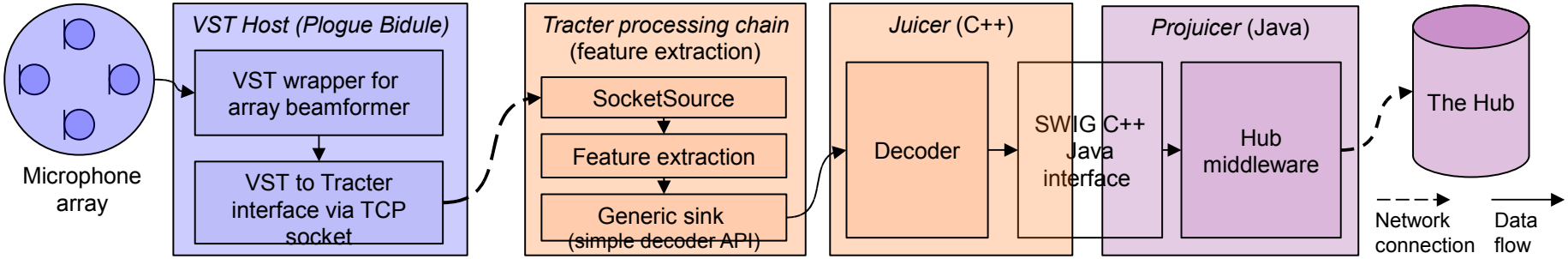
	AMIDA (Idiap)	ICSI-SRI
IHM	27.4	25.5
IHM (ref)	23.5	23.8
MDM	33.2	33.3

Key Achievements
Automatic Speech Recognition

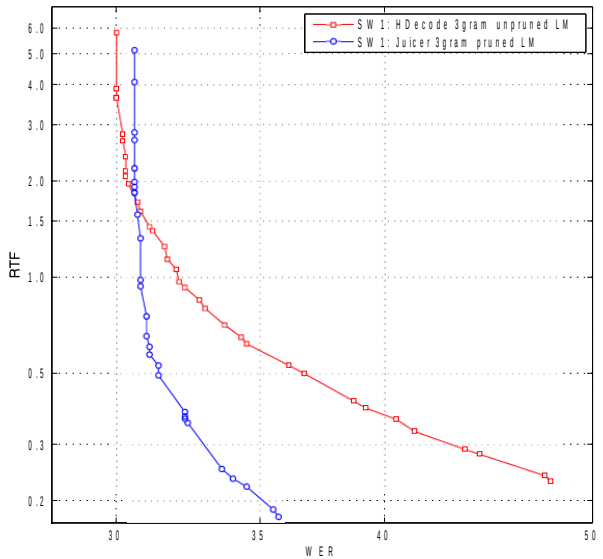
- Real-time, online meeting room ASR
  - Offline systems
    - Big, complicated and precariously glued together by half a dozen different scripting languages
  - Online systems
    - Need to be reliable, fast, low latency, have limited resources (minimal parallelization), all while making least possible sacrifice to accuracy

Key Achievements
Automatic Speech Recognition

- Real-time, online meeting room ASR



- RT-ASR data flow framework (above)
- Juicer decoder performance (right)



Key Achievements
Automatic Speech Recognition

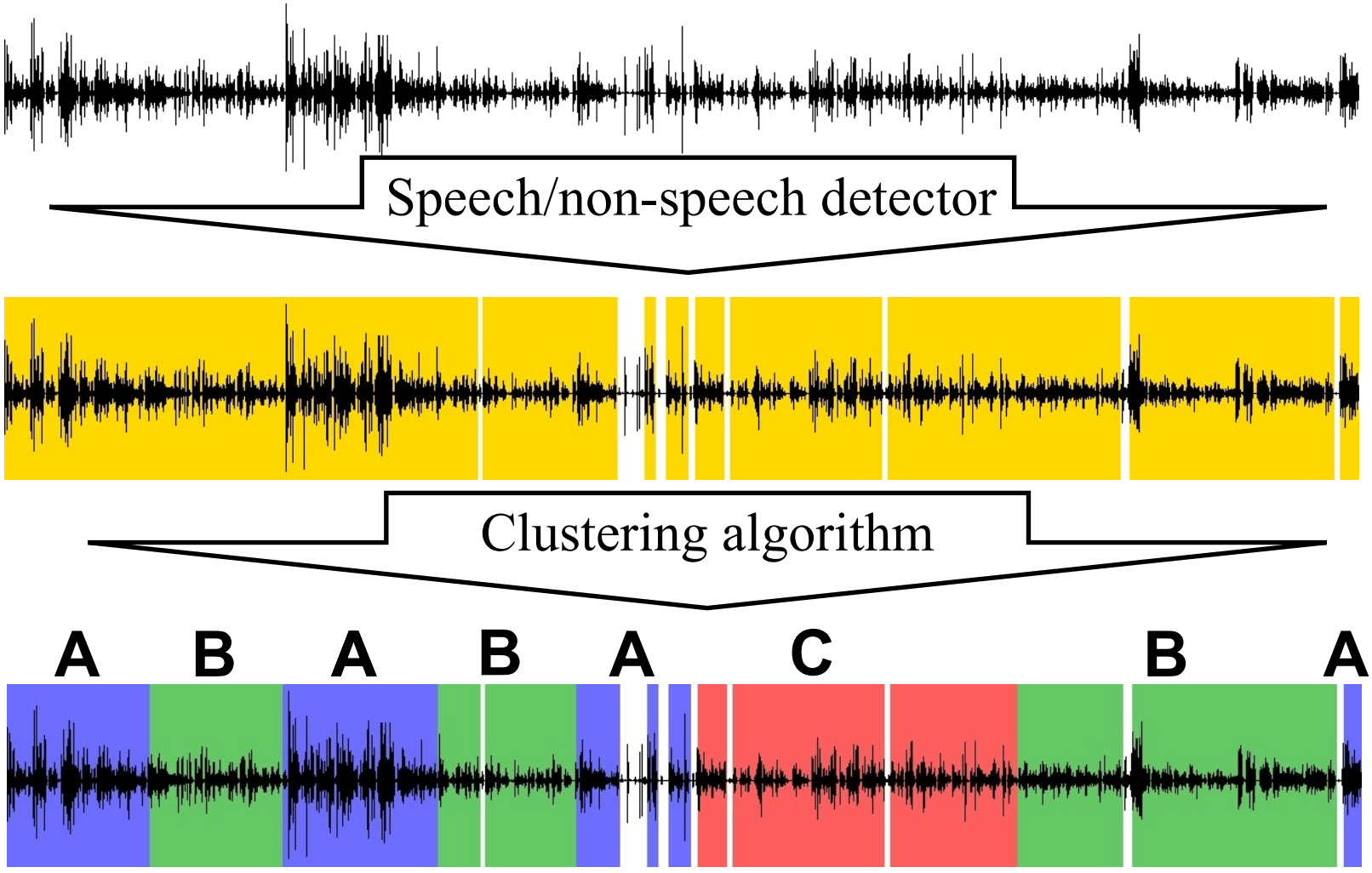
- Publicly released software
  - "Juicer" ASR decoder:
    http://www.idiap.ch/software/juicer
  - "Tracter" ASR front-end:

    http://www.idiap.ch/software/tracter
  - "Beamformit" microphone array beamformer:

    http://www.icsi.berkeley.edu/~xanguera/beamformit

Key Achievements

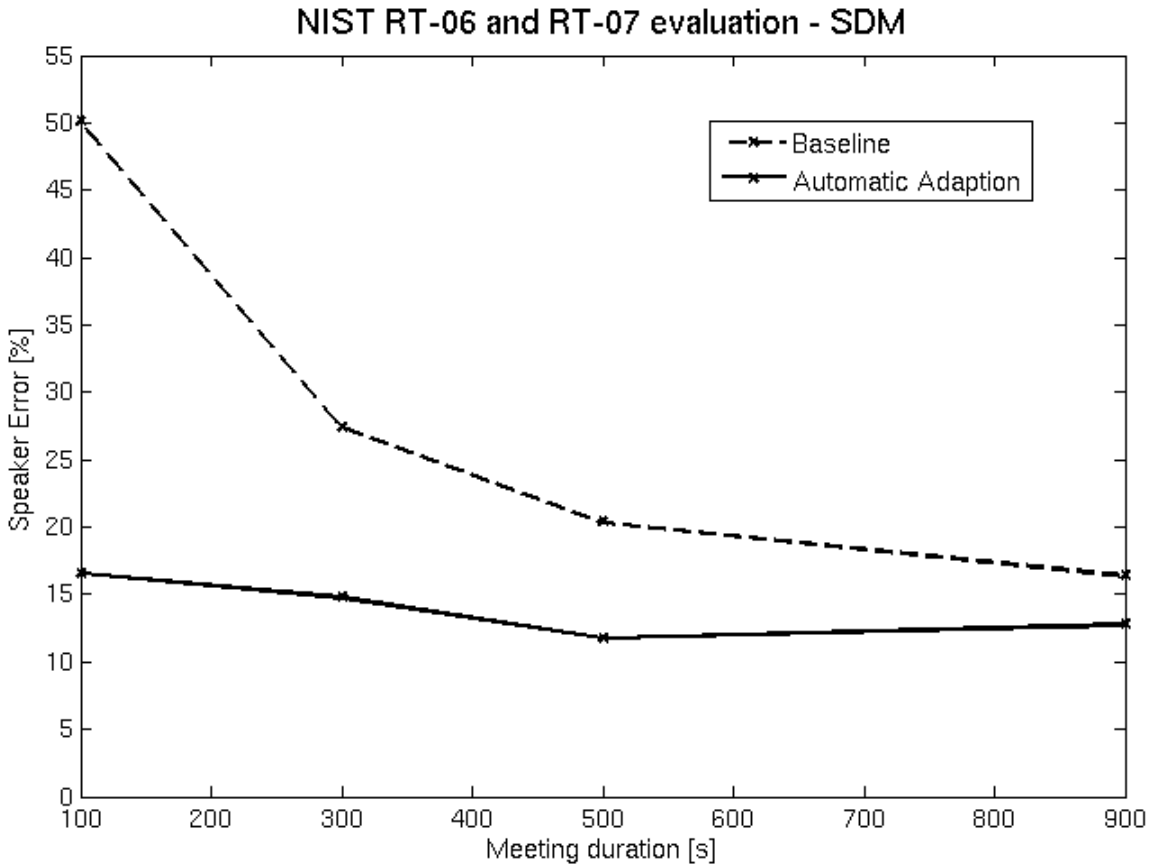
Diarization



Key Achievements
Diarization

Short Speaker Diarization:

Automatic adaption of the key parameters to the meeting length



Key Achievements
Diarization

- Agglomerative Information Bottleneck (AIB)
  - Non-parametric clustering
    - X is a set of elements to cluster into set of clusters C and Y be set of variables of interest
      - Relevance variables are generated by a background GMM with components Y
    - Cluster representation C should preserve as much information as possible about Y while minimizing the distortion between C and X
    - Objective is to minimise

$$I(Y,C) - \beta I(X,C)$$

Key Achievements
Diarization

- Recent investigation of stream combination strategies
  - Typically, combination of features with diverse statistics is addressed by use of ad hoc stream weights
  - AIB enables combination of relevance variable postr distributions

$$P(Y|X) = \sum_i P(Y|X, M_{F_i}) P_F^i$$

where M_{F_i} is background GMM for feature F_i

P_F^i is prior probability of stream F_i

Key Achievements
Diarization

- Realignment procedure using IB criterion
  - Refine speaker boundaries by realigning our cluster models with underlying features
  - Maximising $I(Y,C)$ equivalent to minimisation of KL-divergence

$$\arg\min_{c} \sum_{t} KL\left(P\left(Y \mid x_t\right) \| P\left(Y \mid c_t\right)\right)$$

  - EM solution for $P(Y|C_t)$ enables realignment on posterior features
  - Simple incorporation of multiple streams

Key Achievements
Diarization

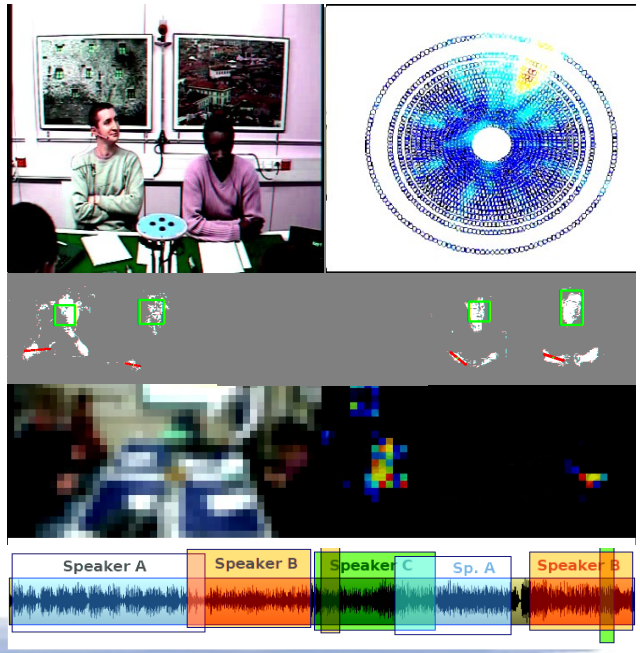
- Evaluation on RT06 Eval

Feature	w/o realign	HMM/ GMM	KL-based
MFCC	19.3	15.7	15.7
TDOA	24.4	25.5	23.9
MFCC+ TDOA	11.6	10.7	9.9

Proposed realignment more effective for multi-stream system

Key Achievements
Diarization

- Audio-visual diarization
  - Initial investigation by ICSI & Idiap showed that visual information (motion, focus of attention) could provide additional robustness to audio-only diarization
  - The price of success!
    - This work is now being conducted in MPR



Key Achievements
Student Theses

- PhD Kofi Boayke (ICSI): "Audio Segmentation for Meetings Speech Processing"
- PhD David Gelbert (ICSI): "Ensemble Feature Selection for Multi-stream Automatic Speech Recognition"
- Masters David Imseng (ICSI-EPFL): "Novel Initialization Methods for Speaker Diarization"
  - Now pursuing PhD at Idiap

Key Achievements
Technology Transfer

- CTI project Veovox
  - Formation of start-up “Veovox” specialising in ASR solutions for order taking systems (e.g. PDA-based restaurant ordering)
  - Small vocabulary, restricted grammar, but extremely robust to noise, accented speech
  - Core technology combines RT-ASR data flow framework with posterior-based template ASR both developed at Idiap
  - First systems should be tested in late 2009/early 2010 with various catering partners

www.veovox.com

Summary

- ASR
  - Good performance in NIST RT evaluation
  - Real-time capability
- Diarisation
  - Optimisation for short meetings
  - Succesful combination of multiple streams
- Tech transfer to VeoVox