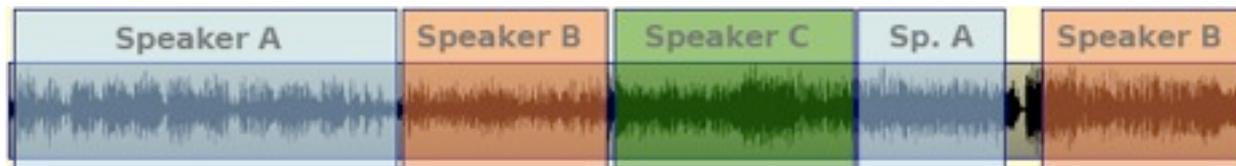
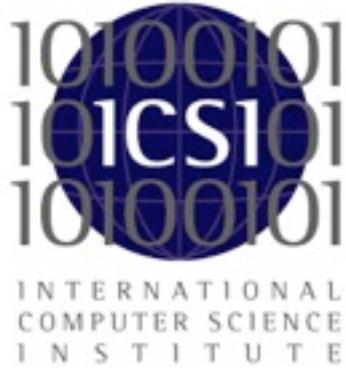


ICSI's Speaker Diarization submissions for RT'09



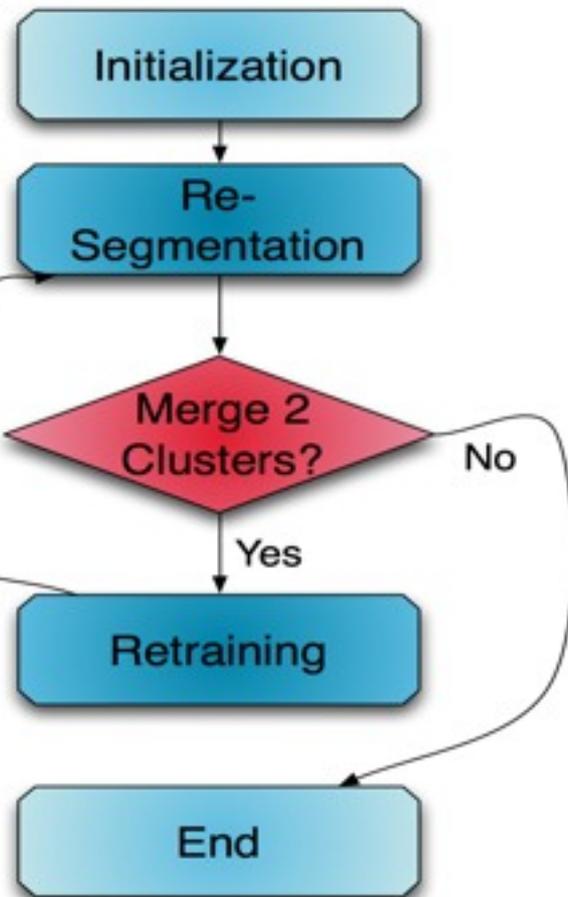
Dr. Gerald Friedland
International Computer Science Institute
Berkeley, CA
friedland@icsi.berkeley.edu



Overview

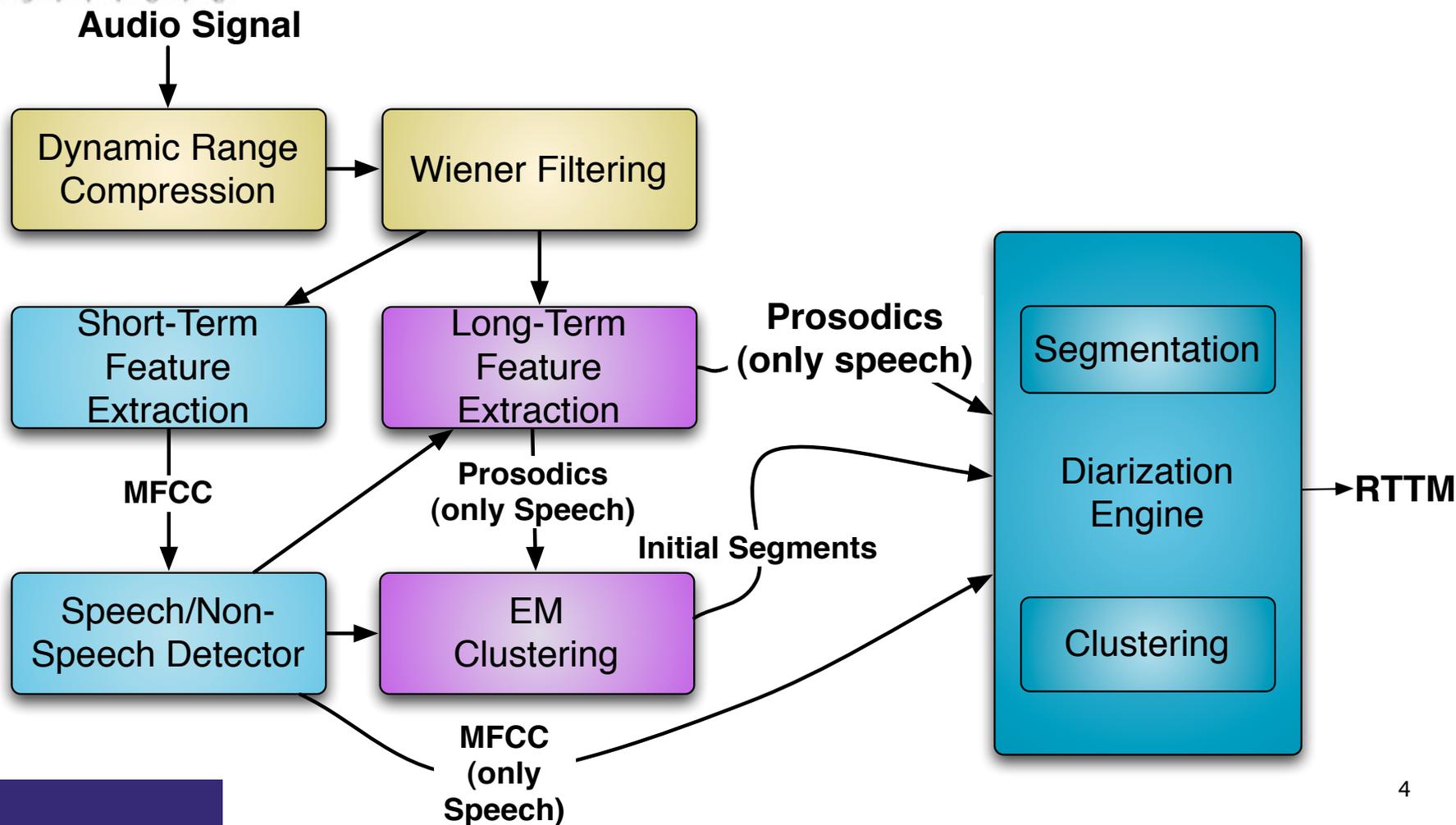
- General overview of the ICSI system
- Tasks submitted: New ideas
- What could we have done better?

Agglomerative Clustering



1. Create k random segments and train k GMMs with g Gaussians
2. Assign frames to clusters according to likelihoods
3. Use BIC to determine if two clusters should be merged.

The ICSI offline diarization system (SDM, only audio)





Reused from ICSI RT'07 System

- Wiener Filter
- Speech/Non-Speech Detector
- Beamformer

Wooters, C. and Huijbregts, M.: The ICSI RT07s Speaker Diarization System, Lecture Notes In Computer Science, Volume 4625, pp. 509--519, Springer-Verlag Berlin, Heidelberg, 2008.

Long-Term Feature Selection

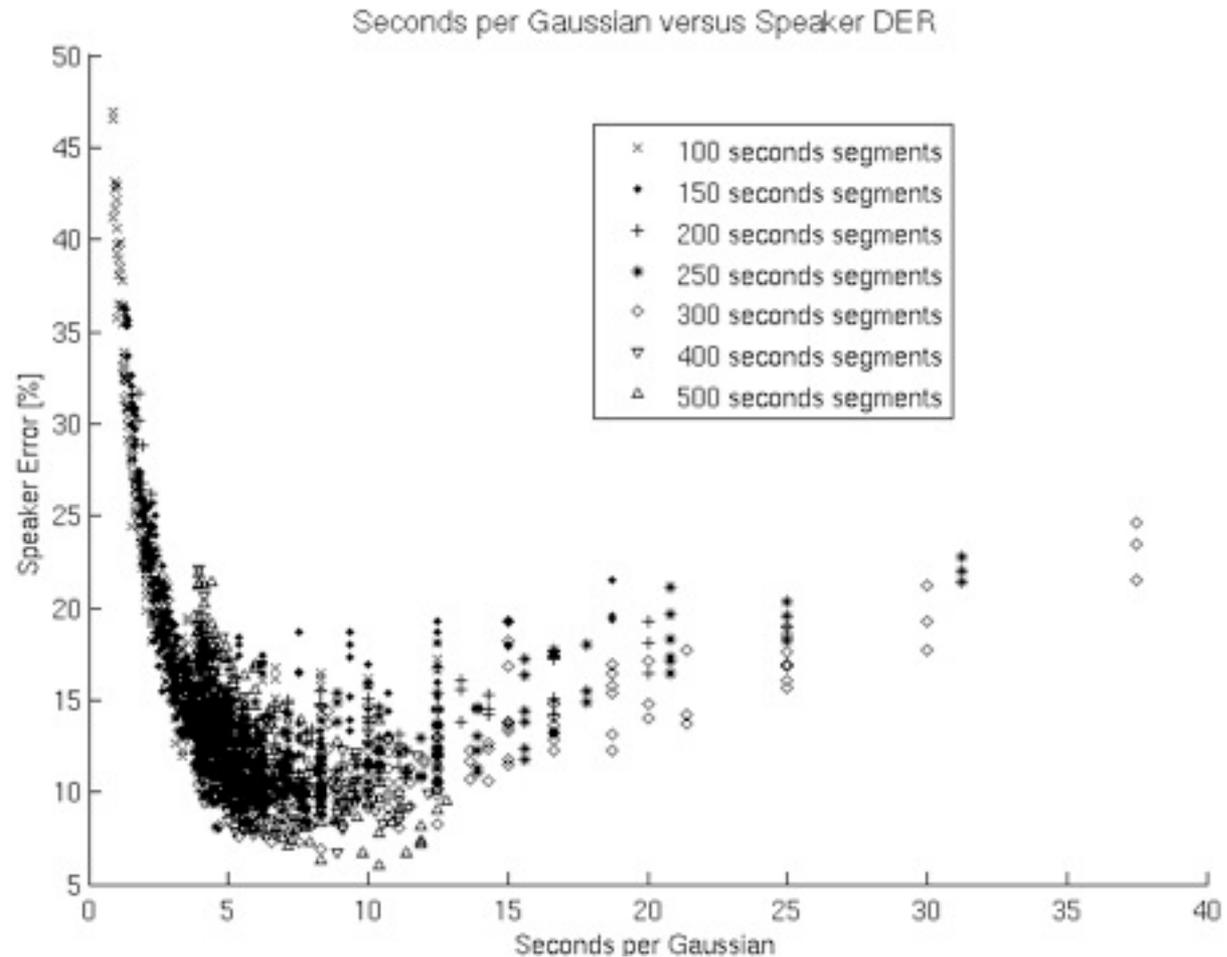
Top 10 of 70 features (500ms):

Prosodic Feature	Intra-Spk Var	Inter-Spk Var	$\frac{Inter}{Intra}$
Pitch Median	17.0	971.2	57.0
Pitch Mean	89.6	1721.9	19.2
F4 Stddev	7.9	56.8	7.2
F4 Min	12.8	80.4	6.2
Pitch Min	28.3	164.6	5.8
LTAS Stddev	90.6	516.4	5.7
F4 Mean	114.9	649.2	5.6
F5 Mean	180.7	929.0	5.1
F5 Stddev	66.7	327.4	4.9
F5 Min	218.3	1032.5	4.7

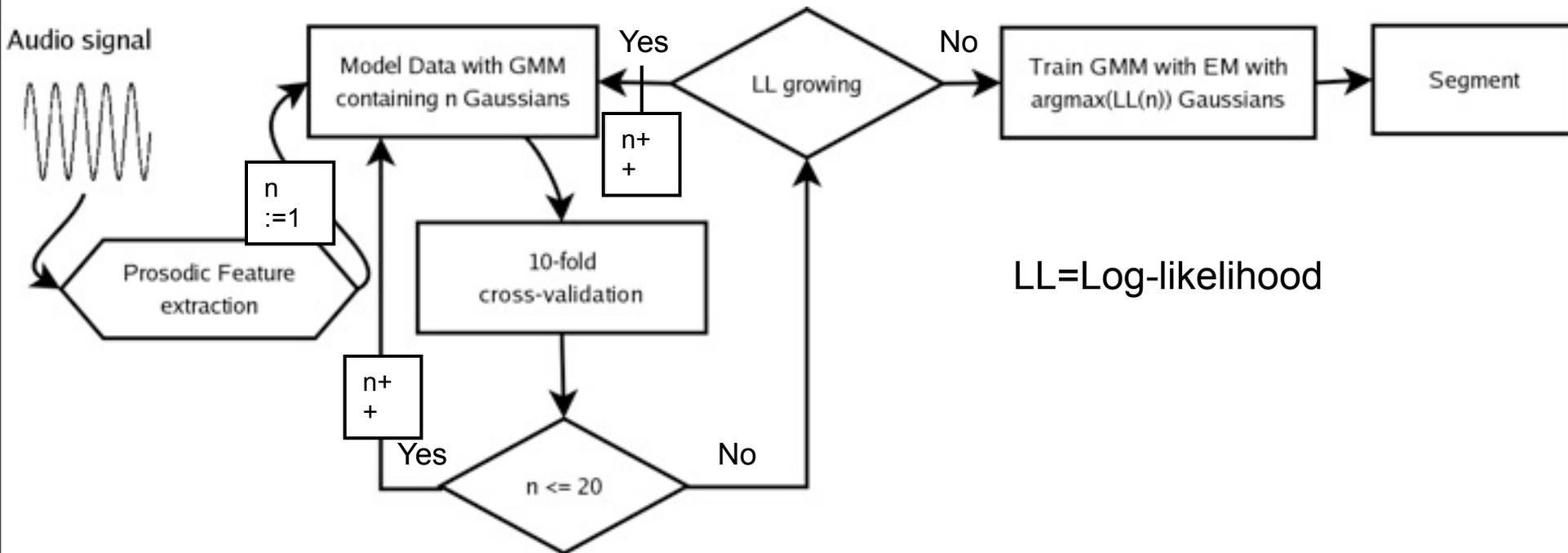
Friedland, G., Vinyals, O., Huang, Y., and Mueller, C.: “Prosodic and other long-term features for speaker diarization”, TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, to appear May 2009.

Parameterless Initialization

Number of Gaussians



Parameterless Initialization



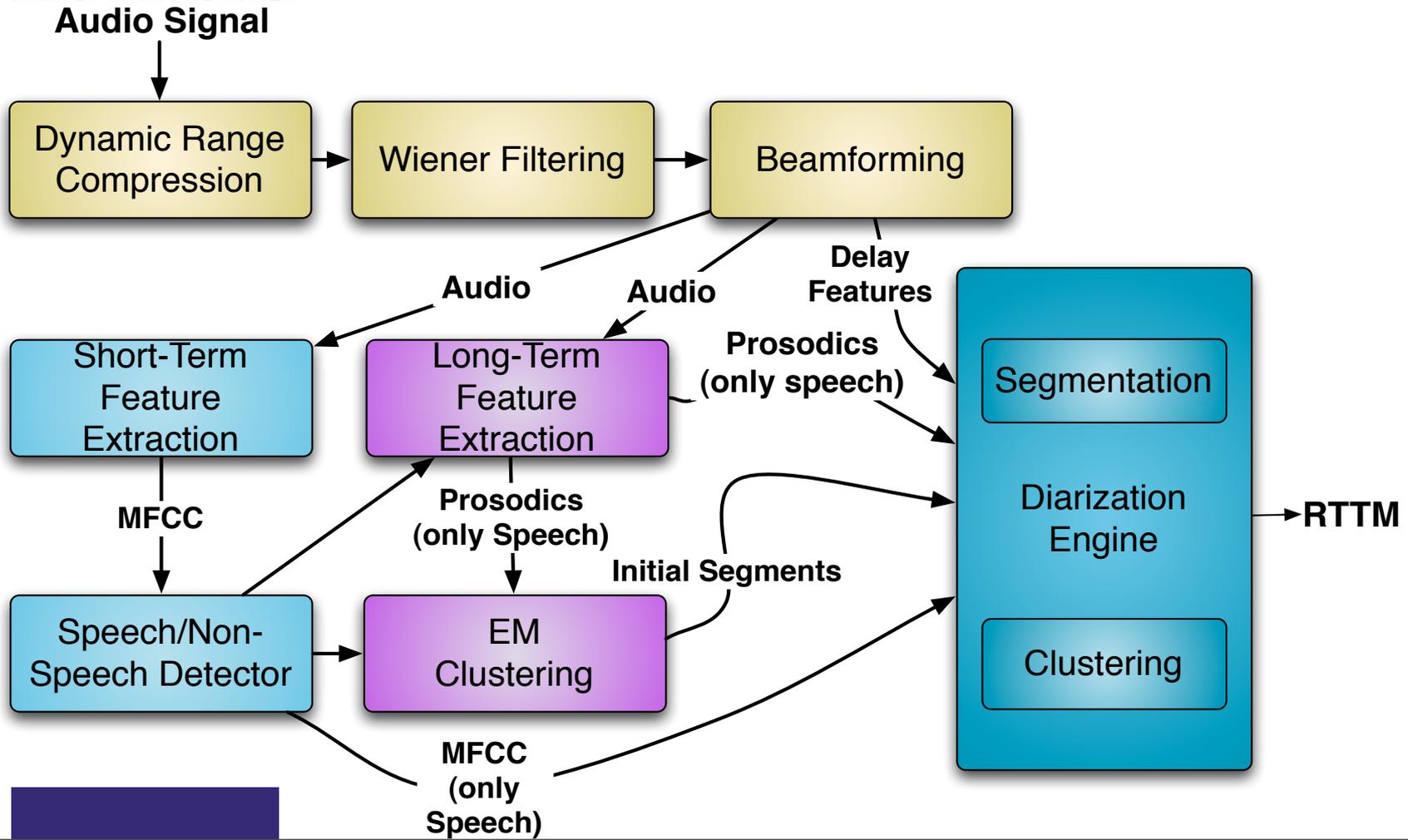
- Number of initial clusters: $\text{argmax}(\text{LL}(n))$
- Non-uniform initialization based on the segmentation at the end

Result?

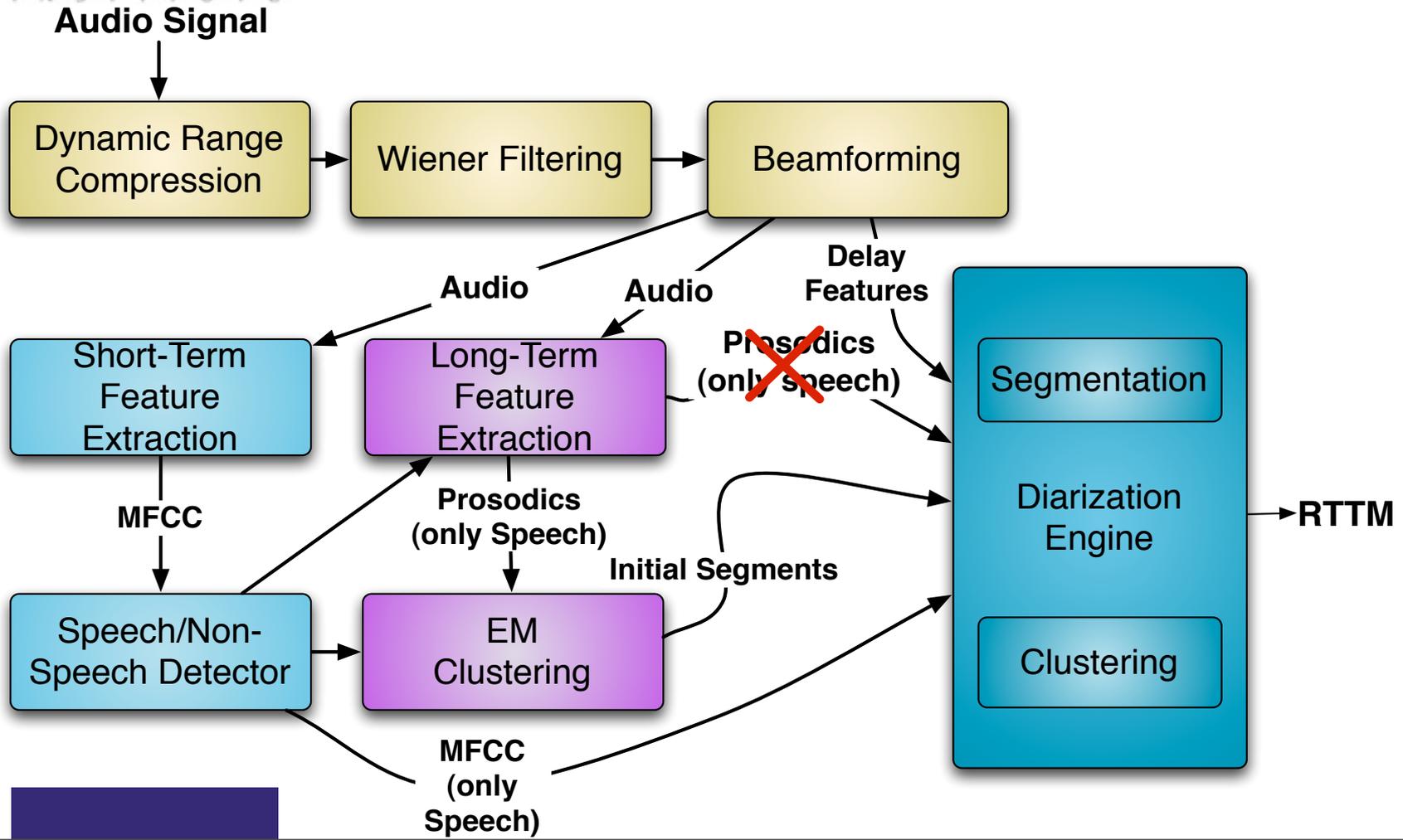
- DevSet '09: 28 meetings from past evaluations (2xAMI, 6xCMU, 4xEDI, 2xICSI, 2xLDC, 6xNIST, 6xVT).

System/Set	DevSet'09 DER
RT'07 SDM System	89.51%
RT'09 Primary SDM Submission	17.00%

MDM Processing Chain (secondary)



MDM Processing Chain (primary)



Result?

System/Set	DevSet'09 DER
RT'07 MDM System	12.2%
RT'09 Primary MDM Submission	9.67%
RT '09 Secondary MDM Submission	9.94%



ADM, MM3A Processing Chain

Same as primary MDM

Low-Latency System (experimental)

- Training:
 - Run primary diarization system (either MDM or SDM) on first 1000 seconds of file.
 - Train GMMs on all speakers and non/speech
- Testing:
 - Classify each 2.5 seconds of the UEM region using the GMMs

Low-Latency System (experimental)

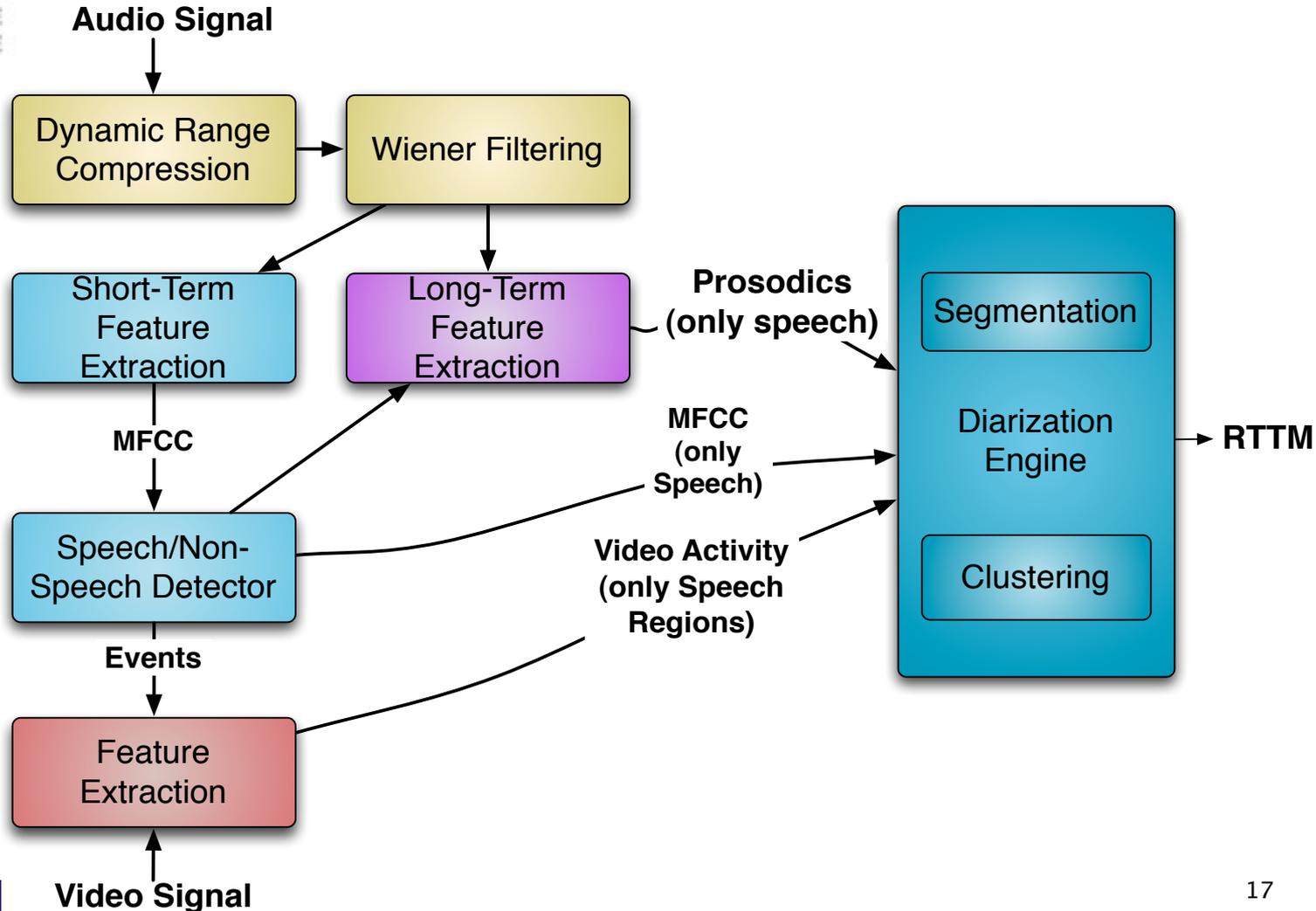
Maximum latency per sample =

$$\max(1000s - t_{UEMstart}, 0) + 2.5s$$

DevSet '09 DER: 31.12% (MDM)

Eval '09 DER: 38.72% (MDM), 44.61% (SDM)

Multimodal Diarization (based on SDM)

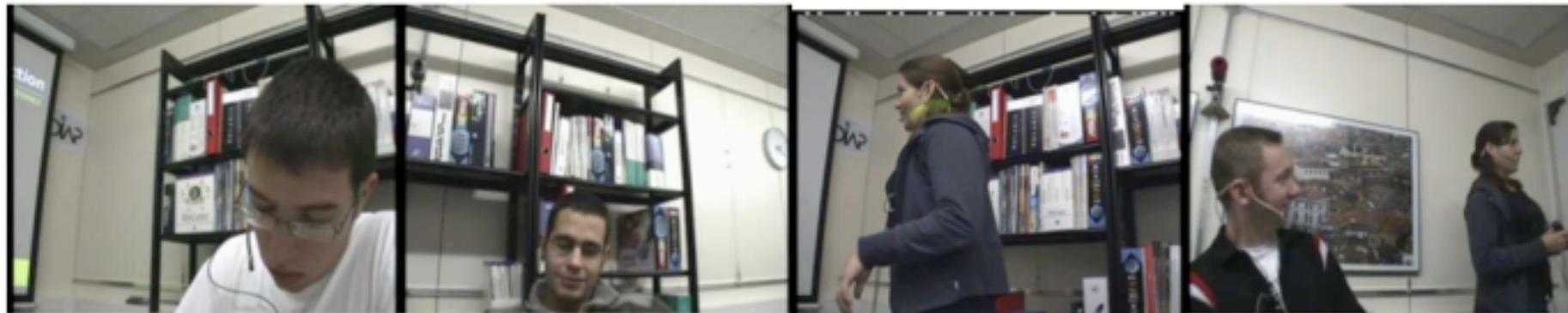




Audio/Visual Correlation Assumptions

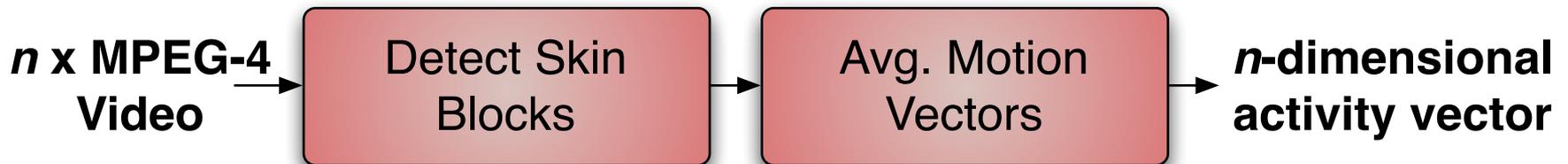
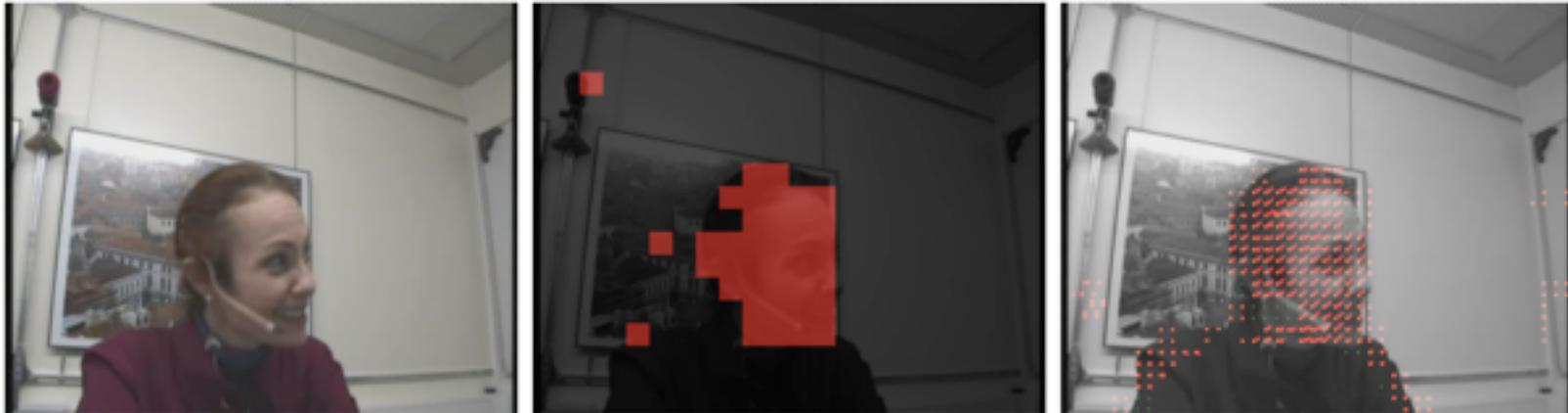
- Camera captures all participants, most of the time.
- Speaker locations have limited spatial variance.
- Speakers have more visual activity than non-speakers.

Video: Different Problems



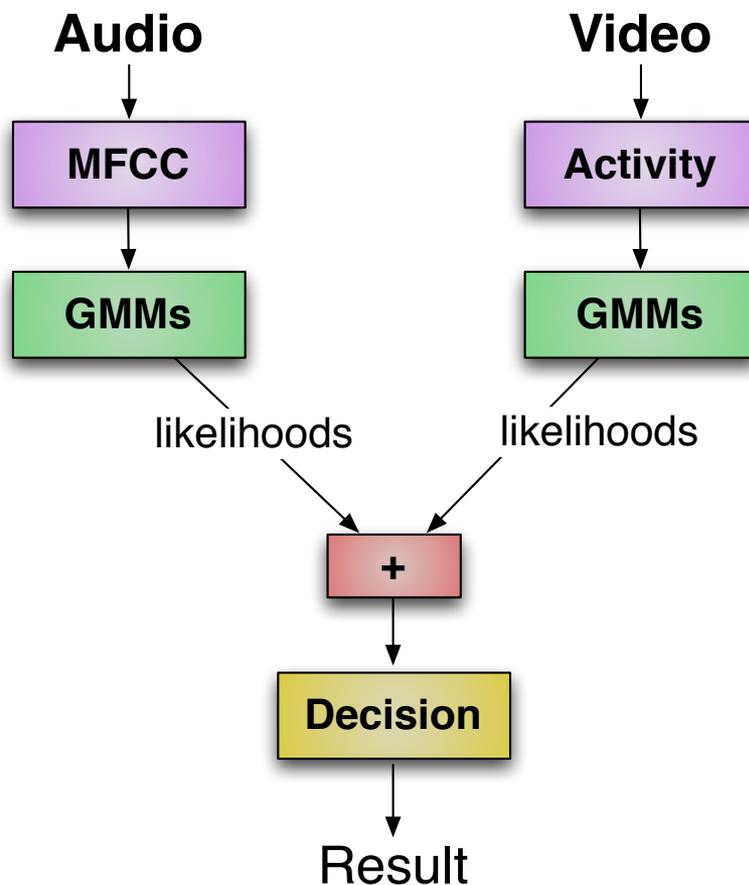
- Close-view still not good enough for face detection
- People lean back and forward, stand up, walk around, leave the room, etc...

Video Feature Extraction



Window size: 400ms

Model-Level Integration



$$\log p(x_{MFCC}, x_{VID} | \theta_i) \doteq (1 - \alpha) \log p(x_{MFCC} | \theta_{i1}) + \alpha \log p(x_{VID} | \theta_{i2})$$

Result?

- Video Development Set:
7 Meetings from Edinburgh,
12 Meetings from IDIAP
(AMI Dataset)

System/DER	VideoDevset '09	Eval '09
RT'07 SDM System	36.34%	?
RT'09 Video System	28.51%	32.56
RT'09 SDM System	?	31.30

Runtime Statistics

Step	Runtime in xRT
Dynamic Range Comp.	0.01
Wiener Filtering	0.48
MFCC Calculation	0.023
Delay Feature Calc.	4.47
Video Feature Calc.	4.78
Prosodic Feature Calc.	7.73
Beamforming	0.55
Prosodic Init	1.083
Speech/Non-Speech	0.806
MS Speaker Diarization	0.571

What could we have done better?

Many things, including:

- Spend more time on engineering:
 - beamformer (better, faster)
 - prosodic feature extraction
- Not trust our alpha parameter

Post-Eval Experiments

System/Set	Eval'09 DER
RT'09 MDM System	17.24%
RT'07 MDM System	16.32%
RT '09 MDM System + inverse entropy alpha guessing	14.53%
RT '09 MDM System + optimally tuned alpha (cheating)	10.70%

Future Work

- (Multimodal) feature integration
- Combine Systems: NTU + ICSI?
- Speech/Non-Speech Detection
- Overlap Detection and Resolution
- Need a better Beamformer (engineering)



Research together with

Research since 2002:

- Past: Chuck Wooters, Yan Huang, Oriol Vinyals, et. al.
- Current: Mary Knox, David Imseng, Adam Janin, Luke Gottlieb, Gerald Friedland
- External: Hayley Hung (IDIAP), Chuohao Yeo (EECS)
- Thanks: Andreas Stolke, Nelson Morgan, Kofi Boakye

Thank You!

Questions?