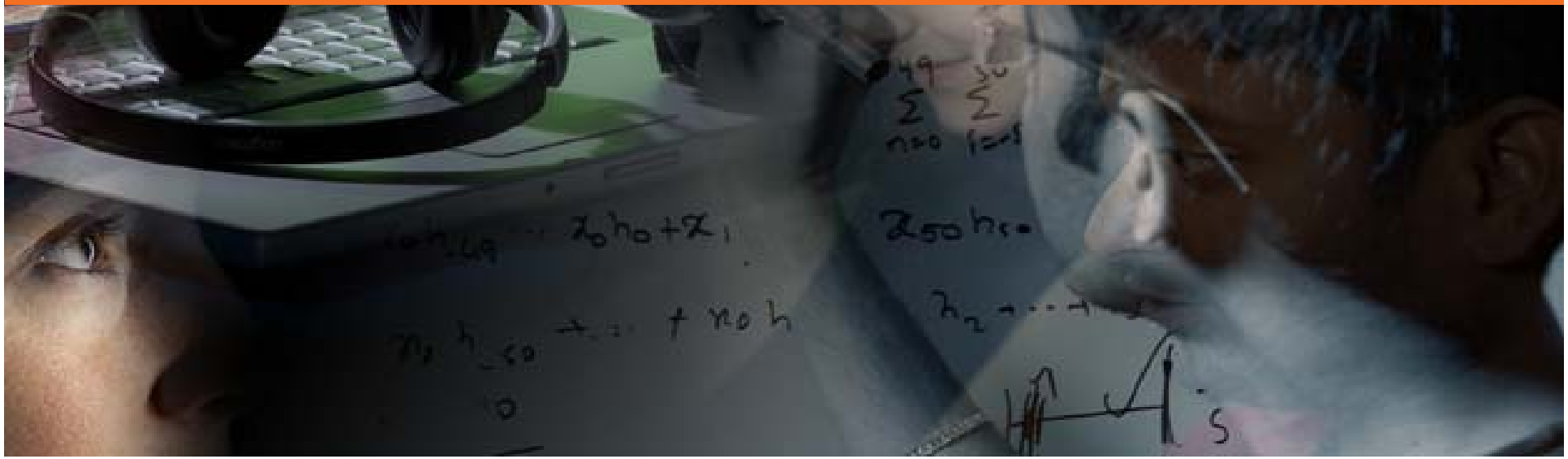
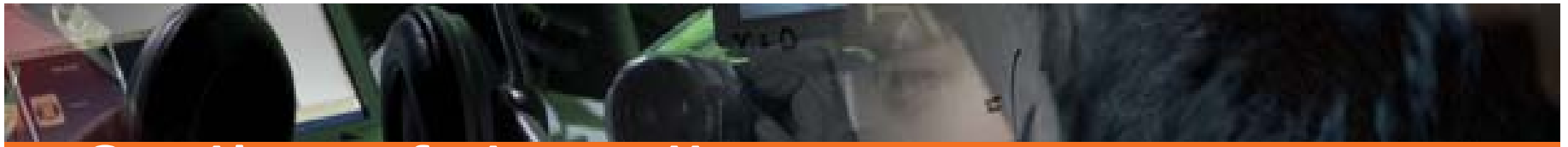
An Information Theoretic Approach to Speaker Diarization of Meetings Recordings

Fabio Valente (joint work with Vijayasenan D. And Bourlard H.)

International Advisory Board Meeting, Idiap, Sep.2, 2011

Research Activities

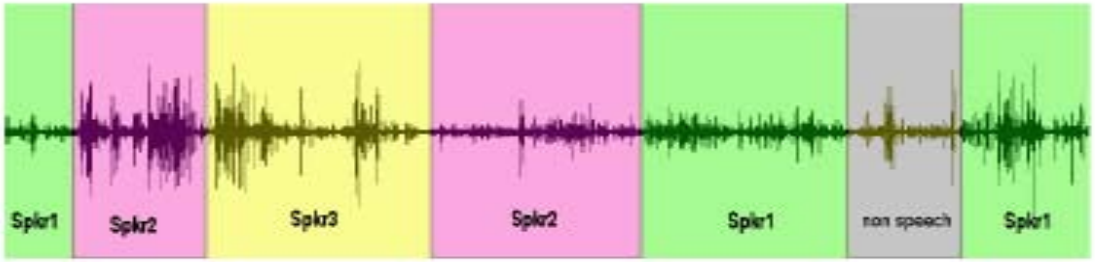




Outline of the talk

1. *Introduction*
2. *Speaker Diarization and Recent Trends*
3. *Information Bottleneck Speaker Diarization*
4. *Towards Manystream Diarization*
5. *Conclusions*

Introduction



Speaker diarization addresses the task of **“who spoke when”**
•Speech/non-speech segmentation
•Estimation of number of speakers
•Identification of speech segments corresponding to each speaker

Applications
•Speaker adaptation in ASR
•Speaker indexing and retrieval
•Conversation analysis

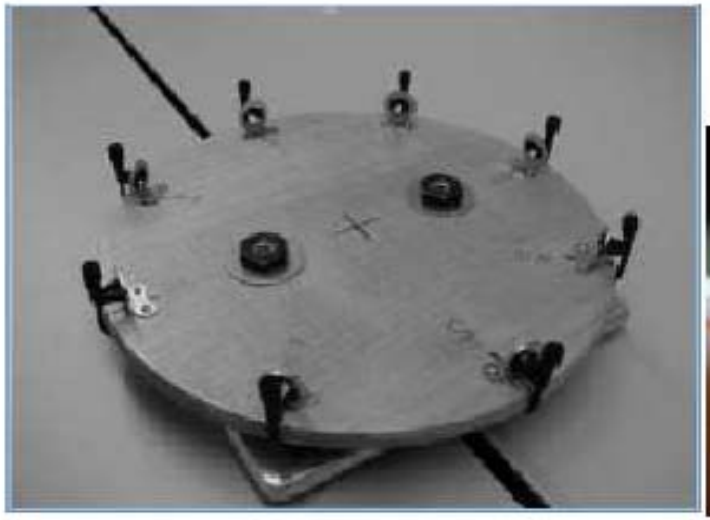
Originally applied to Broadcast News data

Meeting recordings

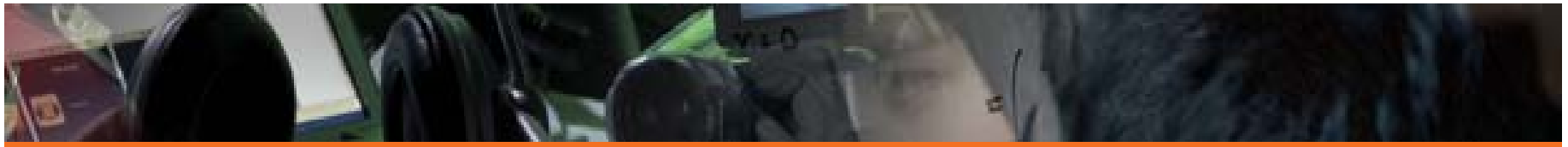
•Smart meeting technologies support recording of face-to-face multi-party conversations a.k.a. meetings

•Further challenges: spontaneous and conversational nature of data

Multiple Distant Microphones

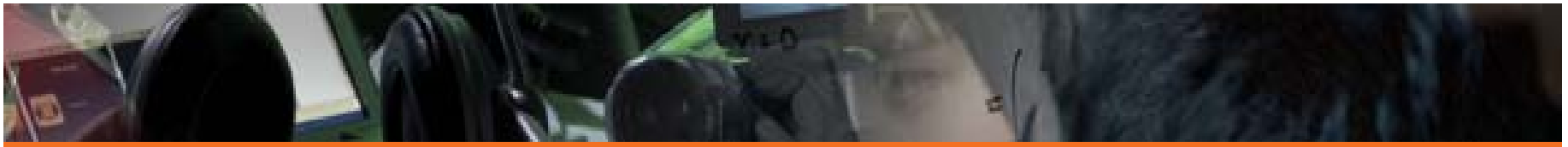


- Non-intrusive audio recording with MDM
- The MDM can consists of microphone arrays or individual distant microphones
- Number and geometry of the MDM vary across different recording environments

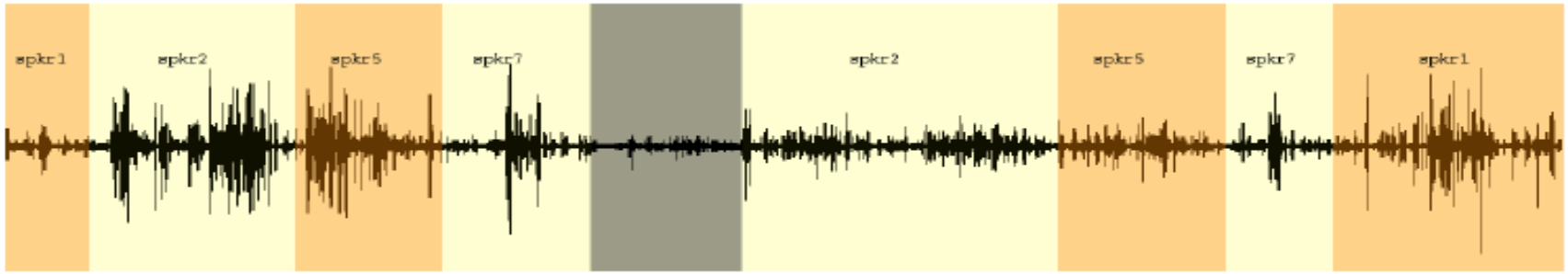


Diarization of Meetings recordings

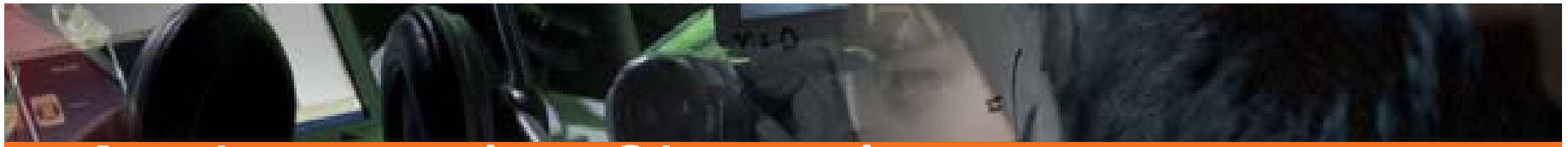
- Since 2005, Speaker Diarization is officially evaluated by NIST in the Rich Transcription campaigns.
- Evaluation dataset consists of recordings across multiple sites :
  - *4-13 speakers*
  - *number of microphones ranges from 2 - 16*
  - *7 different meeting rooms*
- Very heterogenous acoustic conditions
- Diarization Error(DER) is used as official metric: sum ofspeech/non-speech error and speaker error.



Conventional BN diarization

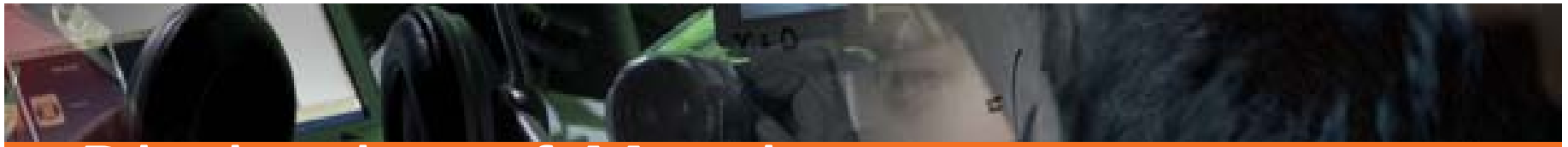


- Short time spectral features (MFCC) as input
- Speech/Non-speech detection
- Speaker change detection
- Agglomerative Clustering using HMM/GMM speaker models with minimum duration
  - Compute a distance measure between all pairs of clusters and merge the closest
  - Viterbi realignment to smooth cluster boundaries
  - Iterates until a stopping criterion is satisfied



Agglomerative Clustering

- Change in Bayesian Information Criterion (BIC) as the merge/stop criterion [Chen&Gopalakrishnan98]
- Trade off between log-likelihood and model complexity used as both as (Distance Measure) and stopping criterion (Model Selection)
- Need parametric model, typically a GMM per each speaker
- Requires estimation of $0.5K(K-1)$ GMMs, where K is the number of clusters
- Computationally expensive as GMM re-estimation require a EM algorithm



Diarization of Meetings

Broadcast News systems evolved towards Meeting recordings in two directions :

1 Several ad-hoc modification to the BIC have been proposed over time to increase robustness and remove dependencies on any tuning [Ajmera et al. 2004]

2 Multiple feature streams have been integrated with MFCC to provide complementary information.
- Prosodic features (+10%)
- Modulation spectrum of the signal (+5%)
- Visual features like head and body movement (+5%)
- Speaker location information form the array (+20%)

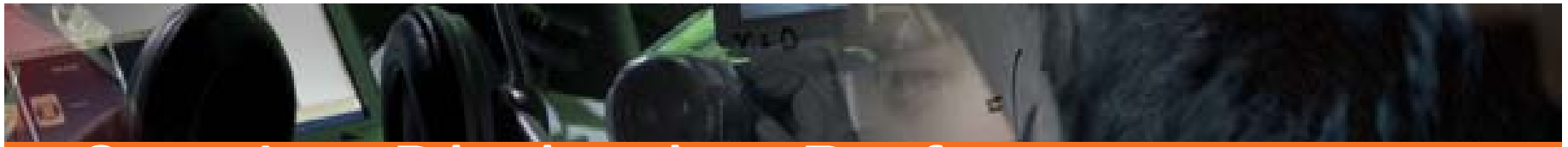
- SoA in diarization consists in MFCC+TDOA combination.

Multistream Diarization

•TDOA carries information on the current speaker location

•Used as complementary features to MFCC in the RT06 evaluation provided the lowest diarization error [Pardo, Anguera and Wooters 2007].

$$\log p(D_j|M_j) = W_{mfcc} \log \left(p(D_j^{mfcc}|M_j^{mfcc}) \right) + W_{tdoa} \log \left(p(D_j^{tdoa}|M_j^{tdoa}) \right)$$



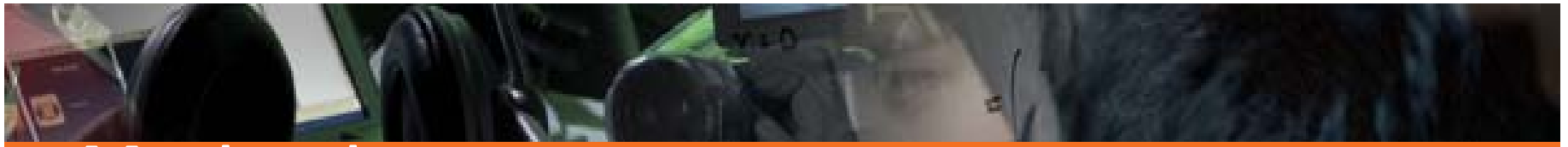
Speaker Diarization Performance

- Speaker Error for RT06 (8 Meetings – 5 rooms)

	MFCC	+TDOA	+MS+FDLP
HMM/GMM	17.0	13.6	15.6

- Real Time Factors

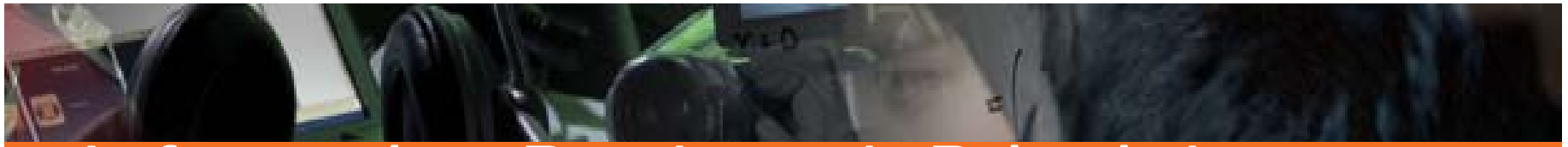
	MFCC	+TDOA	+MS+FDLP
HMM/GMM	3.5	5.8	11.3



Motivations

- Allow the inclusion of many features with very different statistics
- Limit the increase in complexity
- Avoid ad-hoc distance definitions and feature selection/compensation

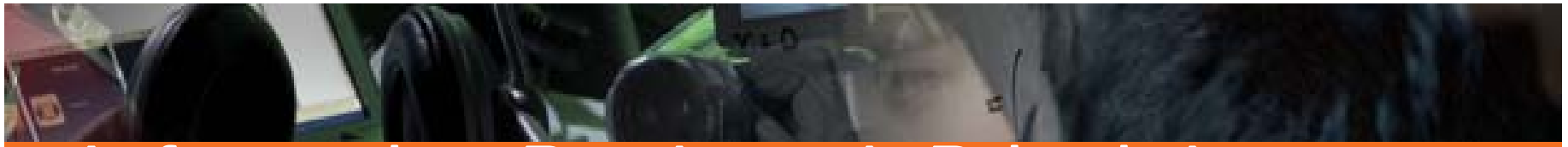
Re-Think entirely the Diarization



Information Bottleneck Principle

- Distributional clustering framework [Thisby and Pereira 2000]
- IB Principle states the optimum clustering C of elements X is the representation that convey as much information as possible about a set of relevance variables Y
- Consider a set of input variables X and associated relevance variables Y . The clustering representation C:
  - maximizes the mutual information with respect to Y i.e., maximizes I (Y , C)
  - is compact i.e., minimize I (C, X)

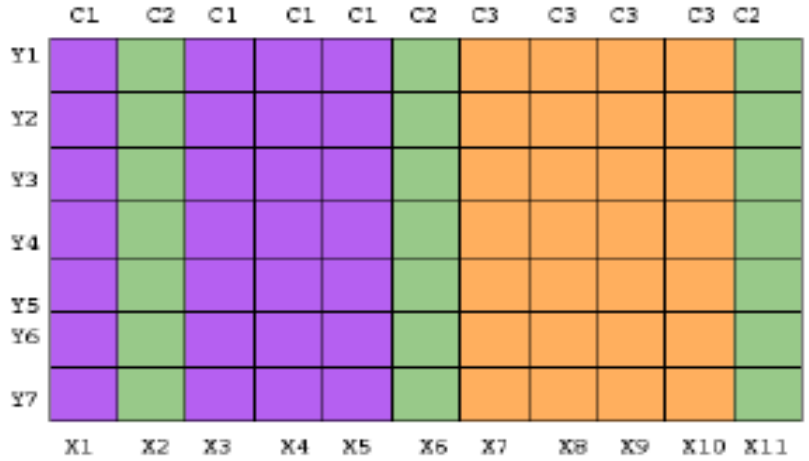
$$\text{Maximize } \mathcal{F} = -I(C,X) + \beta I(Y,C)$$



Information Bottleneck Principle

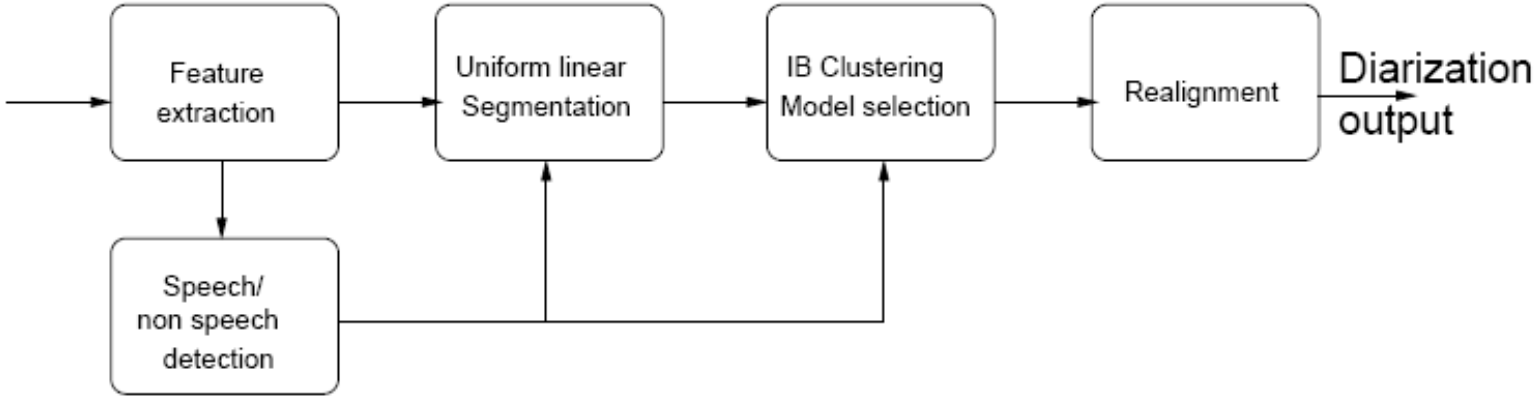
- Relevance variables Y are variables that carry information about the problem
- Example – Document Clustering
  - Similar documents may contain similar words
  - Vocabulary of words can be used as relevance variables
  - Documents that contain similar vocabulary will be clustered together
- Pass the original I (Y ,X) that depends on P(Y |X) throught the bottleneck C
- Merge the two elements X1 and X2 that produce the smallest mutual information drop
- The drop in the objective function can be obtained in **closed form** as sum of KL divergences

Agglomerative IB

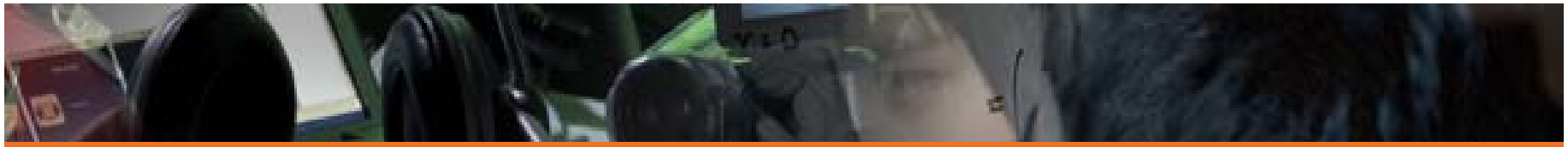


- Estimate P(Y |X)
- Initialized with every element of X as a singleton cluster
- Two clusters X1 and X2 that result in the minimum loss of IB function are merged
- The loss in merging two clusters is obtained in **closed form (JS divergence)**
- The new ”model” P(Y |C) is build simply **averaging P(Y |X1) and P(Y |X2)**
- Continue until a stopping criterion is met

IB Speaker Diarization

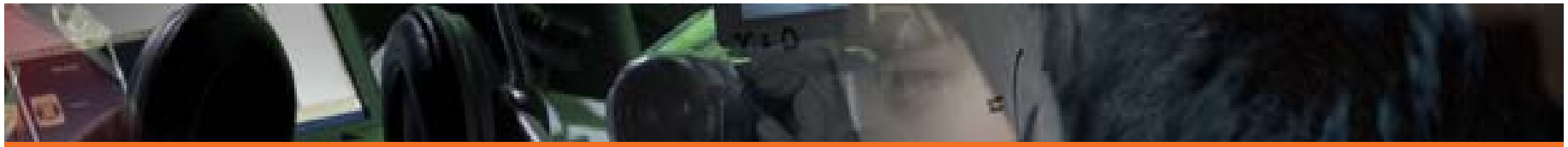


- Input X : speaker homogenous segments
- Relevance Variables Y: components of a background GMM
- Relevance variable distribution p(Y|X)
- The entire system operates based on **p(Y |X)**



Comparison with HMM/GMM clustering

	HMM/GMM	IB
Modeling	A GMM per speaker	Variables Y : p(Y|X)
Distance	BIC and modified versions	JS divergence
Model Estimation	EM	Averaging P(Y|X)



Comparison with HMM/GMM clustering

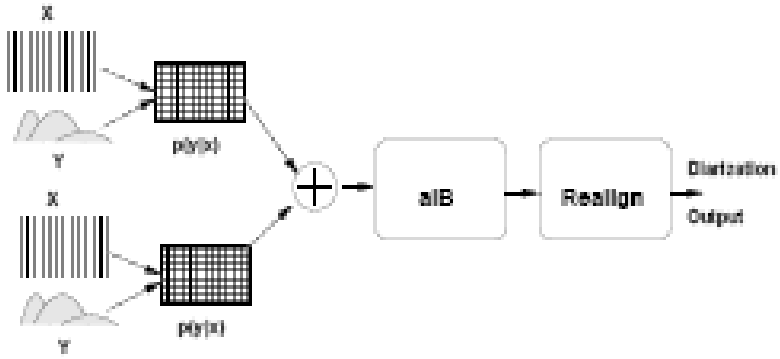
Speaker Error

HMM/GMM	AIB
17.0	17.1

Real Time Factor

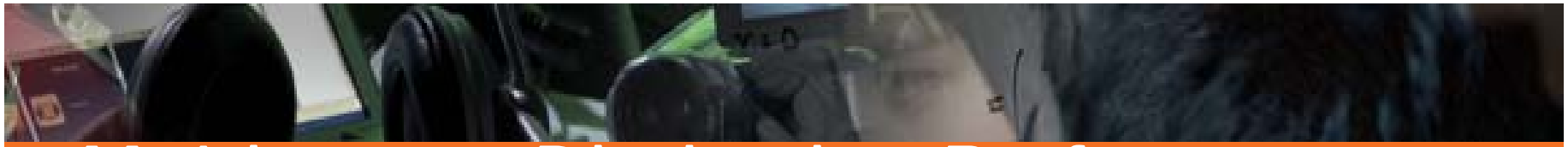
	Estimate p(Y|X)	IB clustering	Viterbi realignment	Total
IB	0.09	0.06	0.07	0.22
HMM/GMM	-	-	-	3.5

Multistream Diarization



- Feature combination is performed in the relevance variable space
- Aligned background GMMs are trained for MFCC and TDOA features
- P(Y|X) is estimated for each features stream using Bayes' rule
- Features combination with weighted linear combination

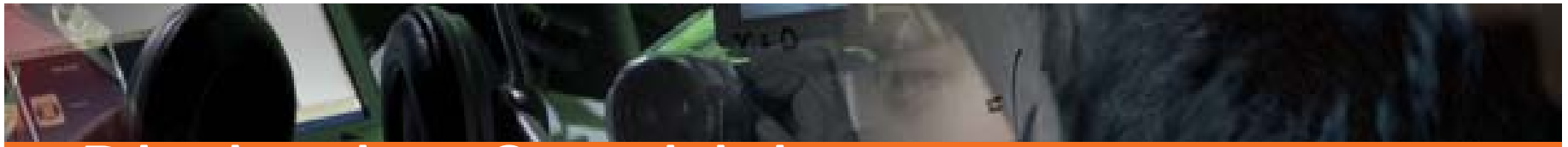
$$P(Y|X_t) = P(Y|X_t^{mfcc})W_{mfcc} + P(Y|X_t^{tdoa})W_{tdoa}$$



Multistream Diarization Performance

	HMM/GMM	IB
MFCC	17.0	17.1
+TDOA	13.6	9.9

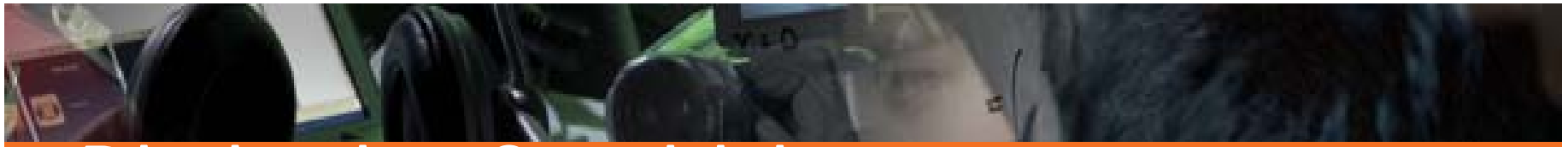
	HMM/ GMM	IB			
		Estimate P(Y|X)	Cluster	Realign	Total
MFCC	3.5	0.09	0.06	0.07	0.22
+TDOA	5.8	0.24	0.08	0.09	0.41



Diarization Sensitivity

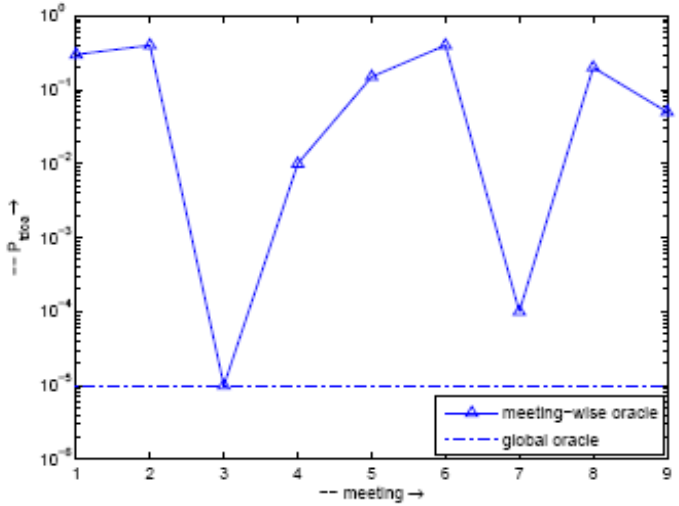
- Evaluation datasets consist of meetings from several recordings environments
- Feature weights are estimated from developement data.

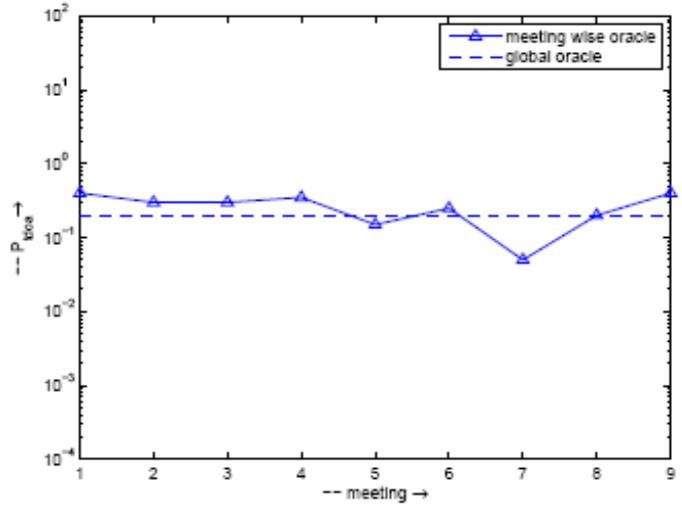
- Oracle experiments to study the combination without any dependencies on development data and provide us with the best possible performance
  - Meeting wise oracle - Set of weights that minimize the speaker error for each meeting.
  - Global oracle weights that minimize the overall speaker error of test data

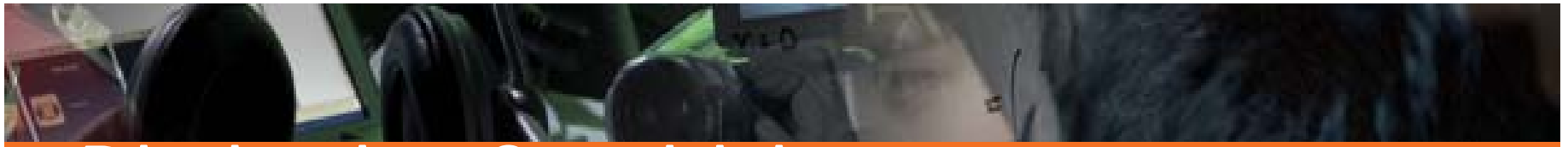


Diarization Sensitivity

	HMM/GMM	IB
Meeting wise oracle	7.0	7.0
Global oracle	11.7	8.7
Estimated weights	13.6	9.9



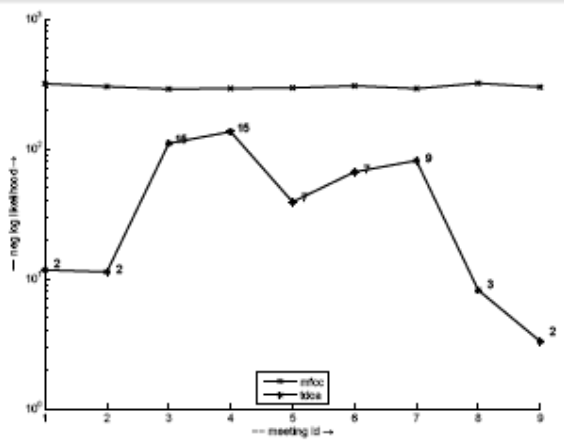




Diarization Sensitivity

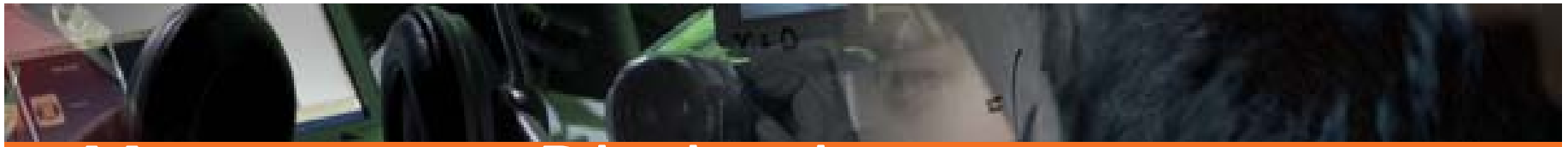
- HMM/GMM combines model log-likelihoods

$$\log p(D_j|M_j) = W_{mfcc} \log \left(p(D_j^{mfcc}|M_j^{mfcc}) \right) + W_{tdoa} \log \left(p(D_j^{tdoa}|M_j^{tdoa}) \right)$$



- IB combines probabilities

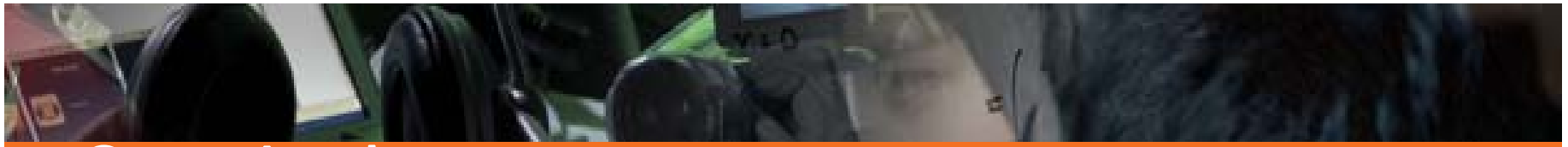
$$P(Y|X_t) = W_{mfcc} p(Y|X_t^{mfcc}) + W_{tdoa} p(Y|X_t^{tdoa})$$



Manystream Diarization

	HMM/GMM	IB
MFCC	17.0	17.1
+TDOA	13.6	9.9
+MS+FDLP	14.5	6.7

	HMM/ GMM	IB			
		Estimate P(Y|X)	Cluster	Realign	Total
MFCC	3.5	0.09	0.06	0.07	0.22
+TDOA	5.8	0.24	0.08	0.09	0.41
+MS+FDLP	11.3	0.52	0.08	0.11	0.75

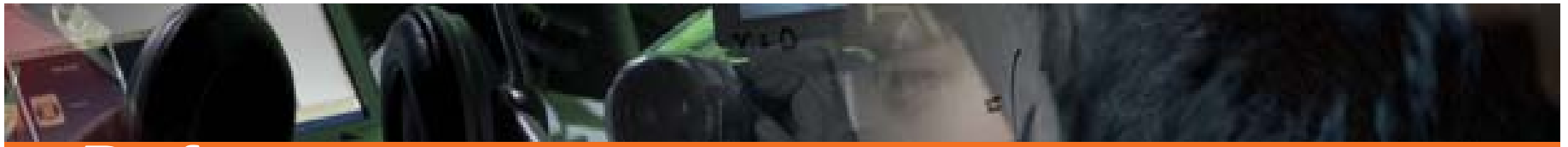


Conclusions

- Allow the inclusion of many features with very different statistics
- Limit the increase in complexity
- Avoid ad-hoc distance definitions and feature selection/compensation

•Non parametric clustering based on Information Bottleneck method
•The system operates entirely in the relevance variable space
•Distance measure arizes from the optimization problem
•Similar performance as in a HMM/GMM system
•More than 10 times faster without any software optimizations for both
•systems

•Multistream diarization performs the combination in the relevance
•variable space
•More robust to variations to the dimensionality of TDOA features
•Limits the complexity of the system
•First success in moving beyond the MFCC+TDOA baseline halfening the speaker error



References

- Vijayasenan Deepu, Valente Fabio and Bourlard Herve *An Information Theoretic Approach to Speaker Diarization of Meetings* Data IEEE Transactions on Audio Speech and Language Processing 17(7) 2009.
- Vijayasenan Deepu, Valente Fabio and Bourlard Herve *An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization* IEEE Transactions on Audio Speech and Language Processing 19(2) 2011.
- Vijayasenan Deepu, Valente Fabio and Bourlard Herve *Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features* Speech Communication (Article in Press).

THANK YOU