Multistream Speaker Diarization through Information Bottleneck System Outputs Combination

Deepu Vijayasenan, Fabio Valente, Petr Motlicek

deepu.vijayasenan,fabio.valente, petr.motlicek}@idiap.ch.

Idiap Research Institute

ICASSP 2011

Introduction and Motivation

- Speaker Diarization determines *who spoke when* in an audio stream.
- In case of meetings data, the recording is done with Multiple Distant Microphones (MDM).
- In case of MDM data, the Time Delay of Arrival (TDOA) of the signal to different microphones can be used as complementary information to acoustic features (e.g. MFCC).
- This combination provides SoA results in Meetings diarization [Pardo2007].

Introduction and Motivation

- Most common combination happens at model level, i.e., a separate model (GMM) for MFCC and TDOA are estimated and then combined by linear weighting [Pardo2007].
- Several studies have discussed the combination of multiple diarization systems:

  [1] Voting schemes between multiple systems.

  [2] Initialization based on diarization output.

  [3] Integrated approaches.
- Can MFCC and TDOA features be integrated using independent diarization systems rather than independent models? Is there any advantage on using systems rather than model combination ?

Introduction and Motivation

- We previously introduced a non-parametric clustering system based on the *Information Bottleneck* principle [Thisby98] working in a space of relevance variables.
- State-of-the-art results using very limited computational complexity.
- Multiple features combination is easily obtained weighting different relevance variable spaces instead of weighting log-likelihoods.
- **Outline of the talk:**

  [1] Information Bottleneck Principle and single stream diarization

  [2] Model based combination

  [3] System based combination

  [4] Hybrid combination

  [5] Experiments

Information Bottleneck principle

- Let X, be a set of elements to be clustered into a set of C clusters.
- Let Y be a set of variables of interest associated with X.
- Let us assume that $\forall x \epsilon X$ and $\forall y \epsilon Y$ the conditional distribution $p(y|x)$ is available.
- IB principle states that the clustering C should preserve as much information as possible between C and Y while minimizing the distortion of C and X.
- This means the following objective function:

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

IB optimization

- Objective function can be optimized in agglomerative or sequential fashion.
- Agglomerative IB [Slonim99]:

  1 Start with trivial clustering of $|X|$ clusters.

  2 Merges clusters that produce the minimum loss in the objective function. The loss can be computed in close form as the Jensen-Shannon divergence.

  3 Merging stops when a stopping criterion is met.

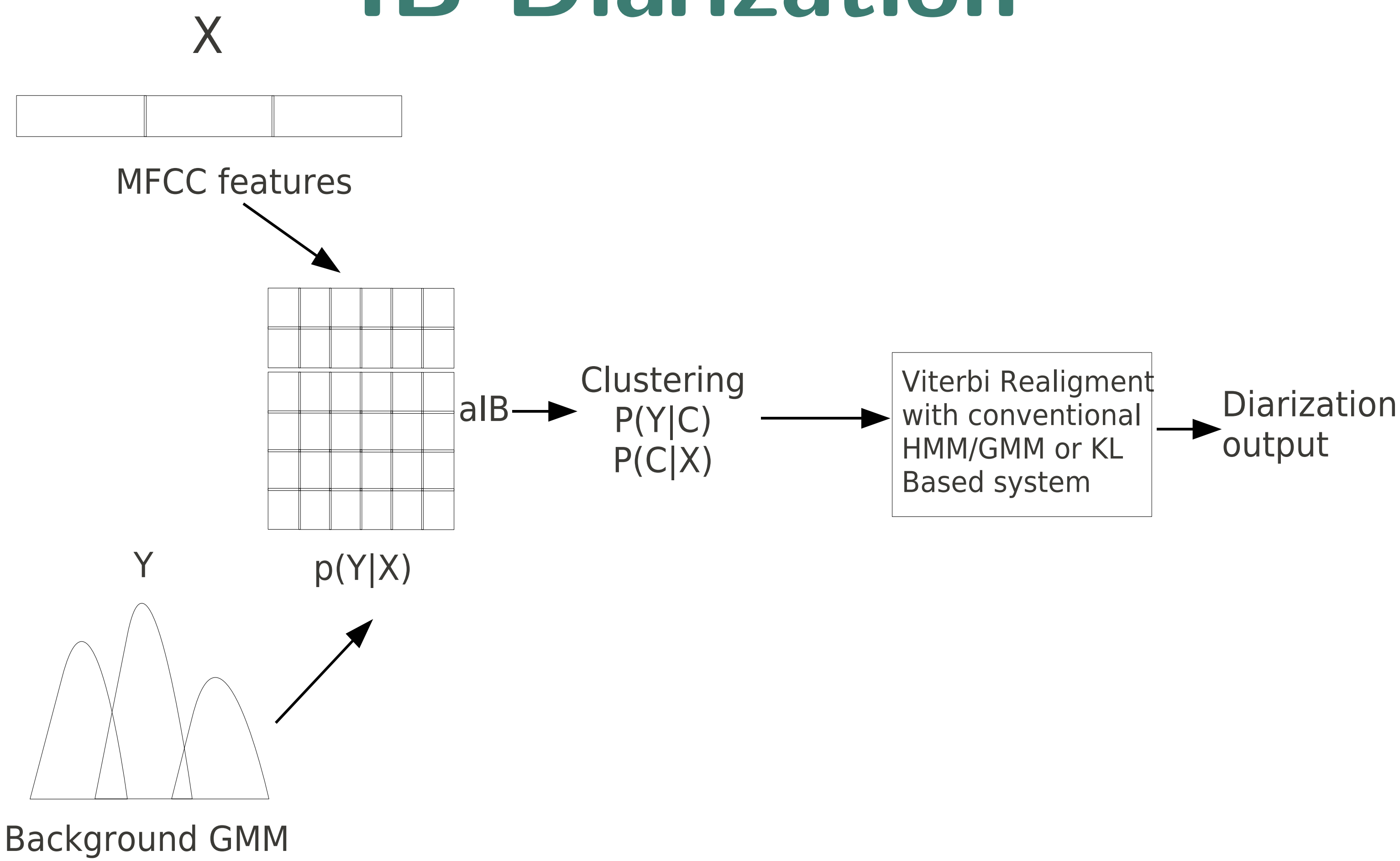
- The output of the aIB is an hard partition of elements $|X|$ in C clusters:

  - $p(c_i|x_t) \in \{0, 1\}$, meaning that each segment is assigned to a cluster (a speaker).

  - $p(Y|C)$ meaning that each cluster is characterized by a relevance variable distributions.

  - The distribution $p(Y|c_i)$ is obtained averaging the distributions $p(Y|x_t)$ for all the segments x_t assigned to the clustering c_i.

IB Diarization



- Elements of X are uniform speech chunks to be clustered.
- Elements of Y are components of a background GMM trained on the entire meeting.
- Probabilities $P(Y|X)$ can be trivially estimated by Bayes rule.

Multiple feature combination

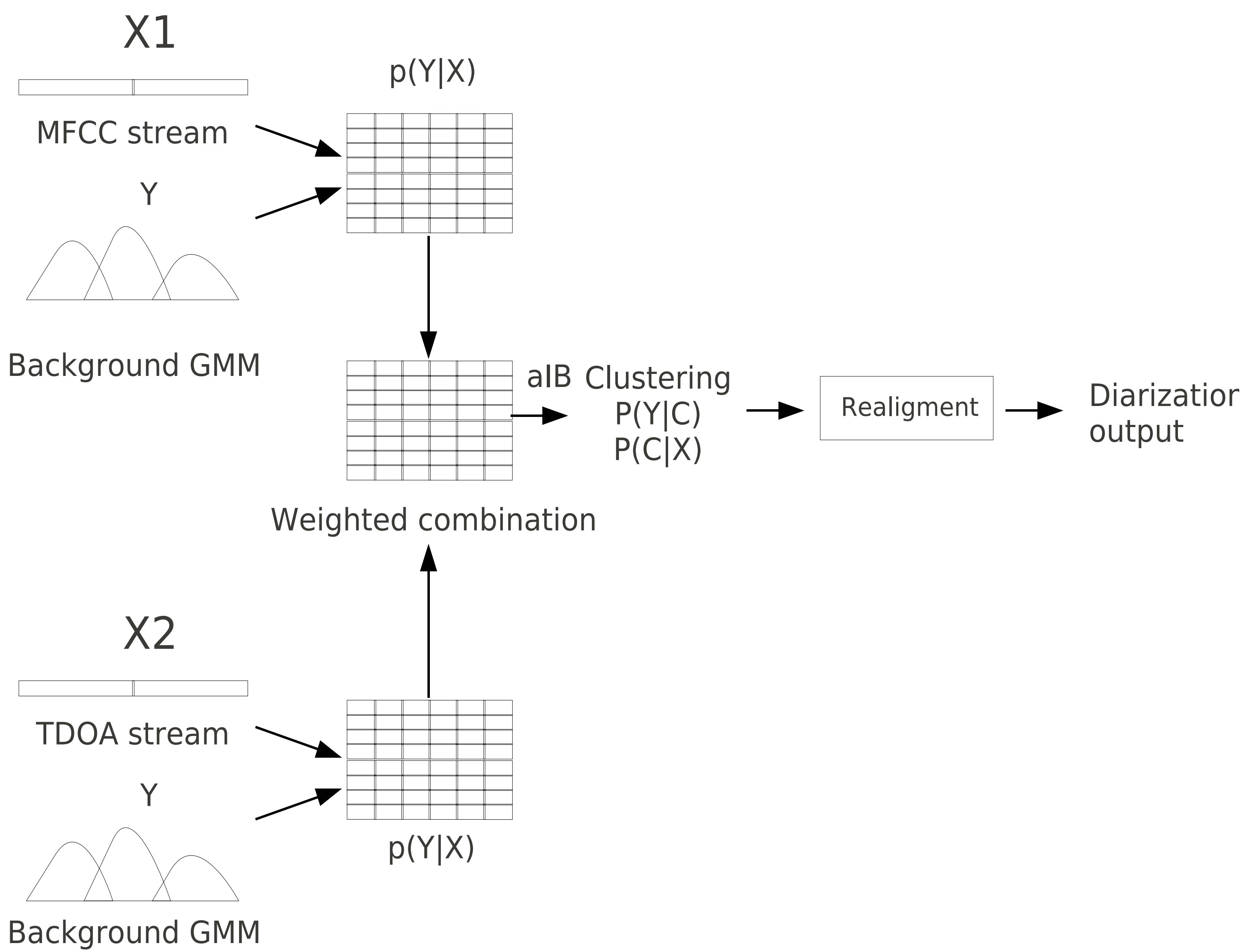
- If MFCC and TDAO features are available, the combination can happen in the space of relevance variables.
- Two aligned background models M_{mfcc} and M_{tdoa} are estimated for each feature stream.
- Two sets of relevance variables $p(Y|x_t, M_{mfcc})$ and $p(Y|x_t, M_{tdoa})$ are then estimated and averaged.

$$p(Y|x_t) = W_{mfcc} \cdot p(Y|x_t, M_{mfcc}) + W_{tdoa} \cdot p(Y|x_t, M_{tdoa})$$

where (W_{mfcc}, W_{tdoa}) are weights and $W_{mfcc} + W_{tdoa} = 1$

- Weights are chosen minimizing the error on the development data set.
- Once $p(Y|x_t)$ is estimated, the rest of the diarization stays the same.

Multiple feature combination



Multiple Systems Combination

- Instead of combining relevance variables before clustering, the weighting can happen after cluster, i.e., *after speaker diarization is performed.*
- Two diarization systems S_{mfcc} and S_{tdoa} based on two aligned background models M_{mfcc} and M_{tdoa}.
- They respectively produce two cluster assignments of segments x_t into clusters c_i: $p(c_i|x_t, S_{mfcc}) \in \{0, 1\}$ and $p(c_i|x_t, S_{tdoa}) \in \{0, 1\}$ as well as two relevance variable distributions for each cluster $p(Y|c_i, S_{mfcc})$ and $p(Y|c_i, S_{tdoa})$.
- Two new distributions of relevance variables $P(Y|x_t)$ can be obtained as:

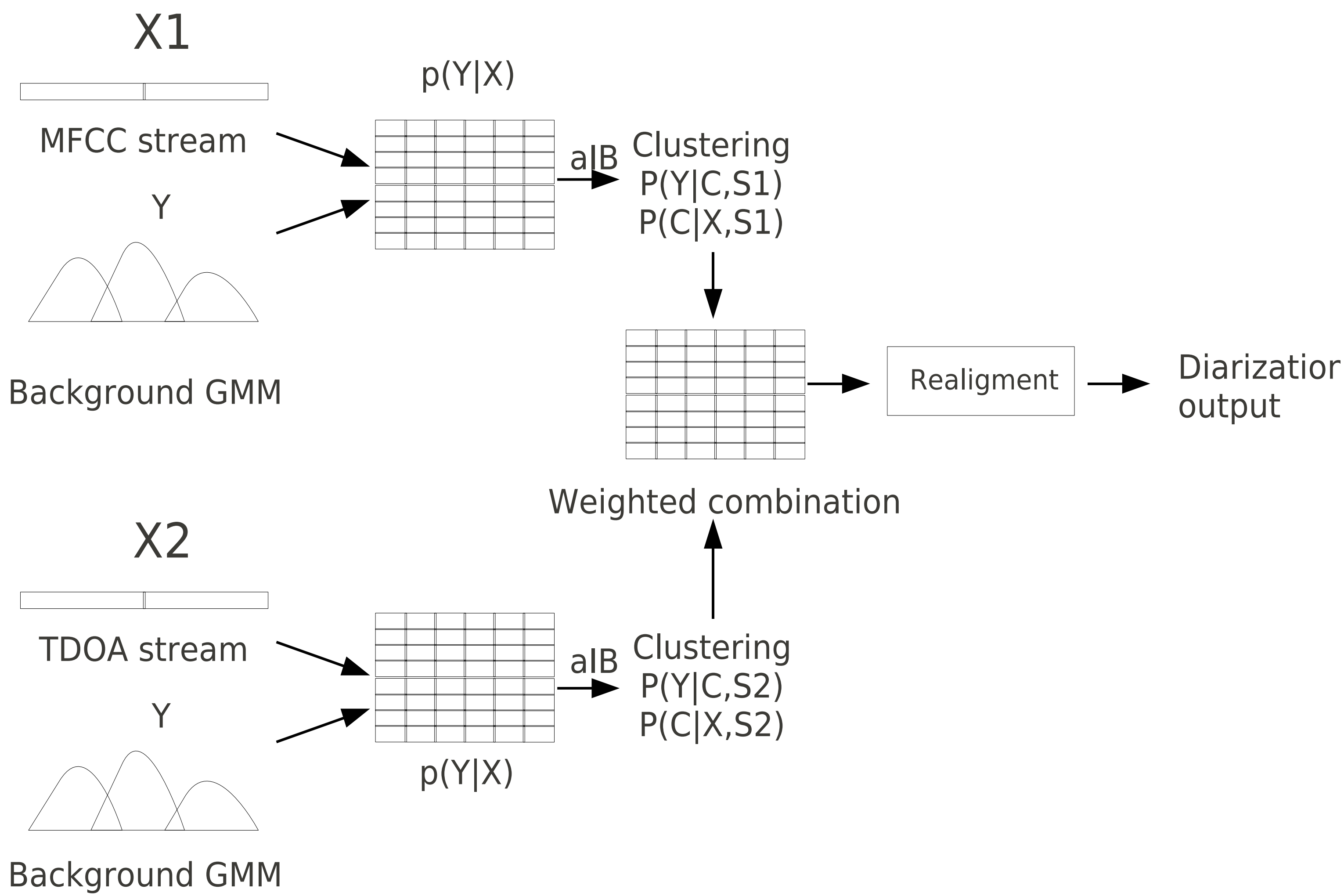
$$P(Y|x_t, S_{mfcc}) = \sum_{c_i} p(Y|c_i, S_{mfcc}) \cdot p(c_i|x_t, S_{mfcc}) \tag{1}$$

$$P(Y|x_t, S_{tdoa}) = \sum_{c_i} p(Y|c_i, S_{tdoa}) \cdot p(c_i|x_t, S_{tdoa}) \tag{2}$$

- the weighting can happen as:

$$p(Y|x_t) = W_{mfcc} P(Y|x_t, S_{mfcc}) + W_{tdoa} P(Y|x_t, S_{tdoa}) \tag{3}$$

Multiple Systems Combination



- Note that $P(Y|x_t, S_.)$ is estimated using all the frames that are assigned to the same cluster thus on significantly more data than in case of model based combination.

Hybrid System-Model Combination

- Instead of combining the relevance variables from two background models or from two diarization systems, a third hybrid solution can be considered.
- It is possible to combine the relevance variable of a first system before clustering with the relevance variables of a second system after clustering, i.e.,

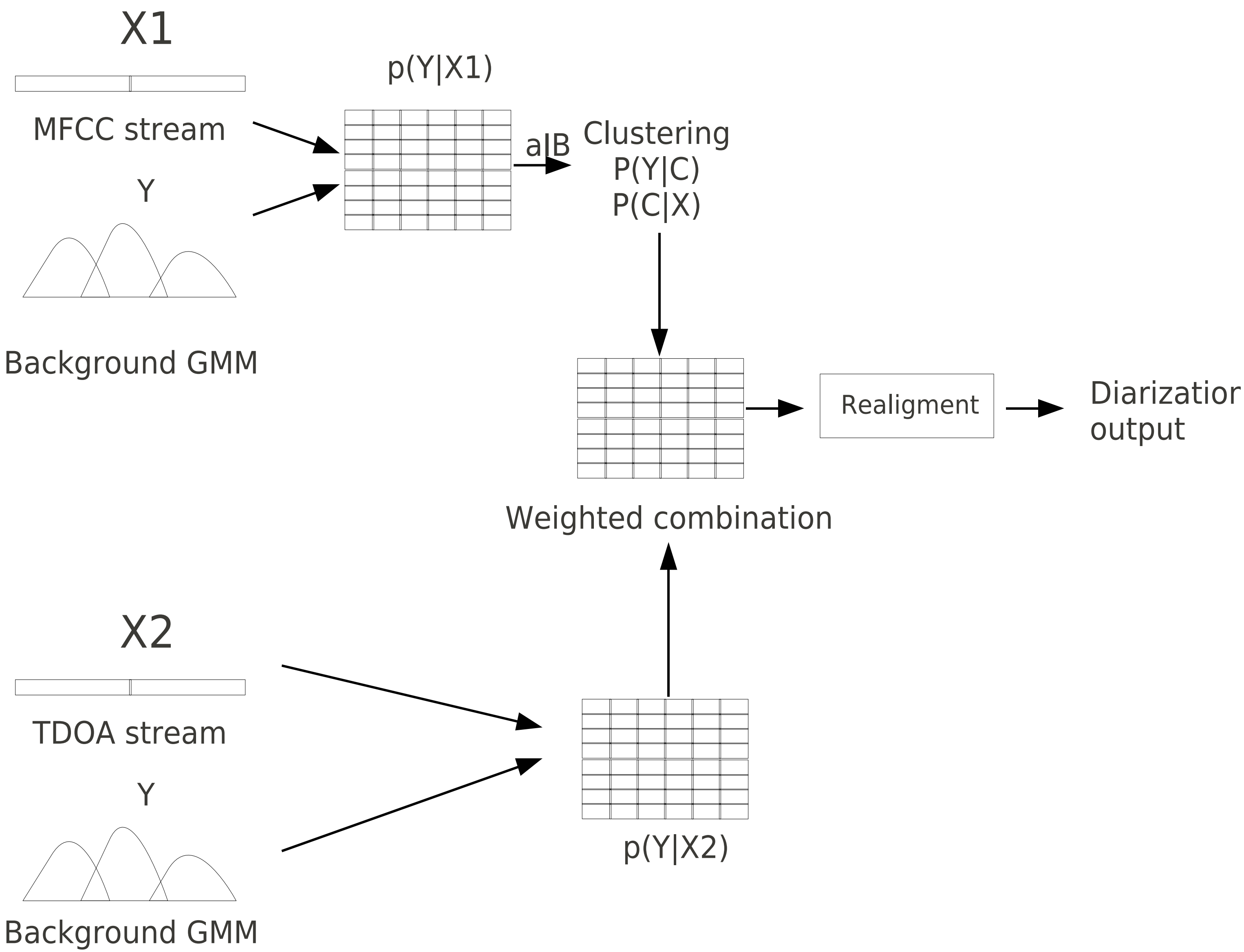
$$p(Y|x_t) \quad = \quad W_{mfcc}\, p(Y|x_t, S_{mfcc}) + W_{tdoa}\, p(Y|x_t, M_{tdoa})$$

a. $p(Y|x_t, S_{mfcc})$ is obtained from the output of a MFCC diarization system.

b. $p(Y|x_t, M_{tdoa})$ is obtained from a TDOA background model.

- In this case $p(Y|x_t, S_{mfcc})$ is estimated using more data than $p(Y|x_t, M_{tdoa})$.
- A similar combination can be obtained inverting the order of MFCC and TDOA.

Hybrid System-Model Combination



Experiments RT

- The experiments are repeated on a collection of 17 meetings from the Rich Transcription (RT) evaluation campaigns.
- Multiple Distant Microphone conditions (MDM), beam-formed to produce a single enhanced speech signal.
- TDAO features are extracted using GCC-PHAT; their dimension is equal to the number of microphone arrays minus one.
- The weights are estimated from a development dataset composed of 12 recordings across 6 meetings rooms.
- The system performance is evaluated using Diarization Error Rate (DER) that is the sum of speech/non-speech segmentation and speaker errors. Since we use the same speech non-speech segmentation across all the experiments only speaker error is reported.

Experiments RT

- Comparison between IB and HMM/GMM whenever MFCC and TDOA features are combined.

	aIB	HMM
Speaker Error	11.6	12.4

- Weights obtained on the development data set are:

	aIB	HMM
(P_{mfcc}, P_{tdoa})	$(0.7, 0.3)$	$(0.9, 0.1)$

- Weighting is different as the two systems combines different quantities: probabilities in case of IB and log-likelihoods in case of HMM/GMM.
- The combination using relevance variables outperforms combination using log-likelihoods.

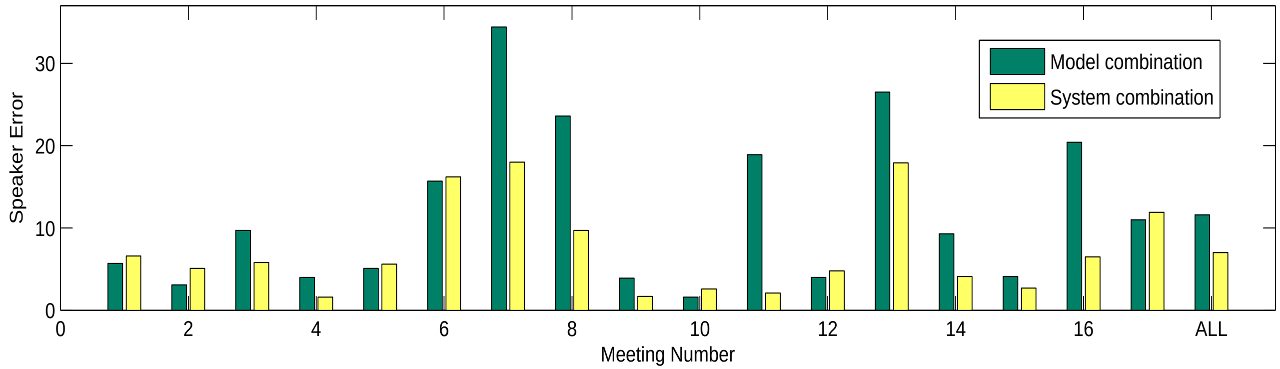
Experiments RT

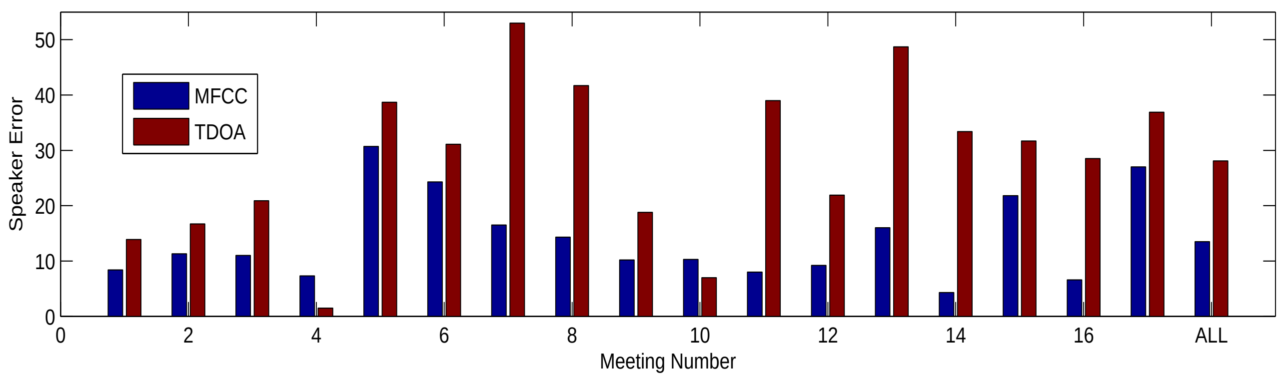
Table 1: Speaker Error for the proposed combination schemes: model based, system based and the two hybrid combinations.

Case	MFCC	TDOA	(W_{mfcc}, W_{tdoa})	Speaker Error
1	Model	Model	(0.7,0.3)	11.6 (–)
2	System	System	(0.7,0.3)	7.3 (+37%)
3	System	Model	(0.8,0.2)	10.5 (+9%)
4	Model	System	(0.6,0.4)	9.4 (+19%)

- System combination largely outperforms other model and hybrid-combinations.
- In the model based combination, $p(Y|x_t)$ is obtained weighting $p(Y|x_t, M_{mfcc})$ and $p(Y|x_t, M_{tdoa})$ estimated using observations from the segment x_t.
- In the system based combination, $p(Y|x_t)$ is obtained weighting $p(Y|x_t, S_{mfcc})$ and $p(Y|x_t, S_{tdoa})$ estimated using the output of systems S_{mfcc} and S_{tdoa} thus significantly more data.

Experiments





- Improvements are larger in meetings where the difference (in terms of speaker error) between MFCC and TDOA is high.
- Weights move towards the feature stream that has been estimated on the diarization output, thus on more data.

Conclusion

- We investigated whether MFCC and TDOA features can be combined trough system based combination.
- The study is based on the Information bottleneck diarization system and three models are proposed:
  [1] Model based combination
  [2] System based combination
  [3] Hybrid model-system combination
- System based combination largely outperforms both model and hybrid schemes.
- Improvement comes from robust estimation of TDOA relevance variables obtained from the diarization output.

Thank You