# Robust Statistical Processing of TDOA Estimates for Distant Speaker Diarization


Pablo Peso Parada[1] - *Nuance Communications*
Dushyant Sharma - *Nuance Communications*
Toon van Waterschoot - *KU Leuven*
Patrick A. Naylor - *Imperial College London*


DREAMS
Dereverberation and Reverberation of Audio, Music, and Speech

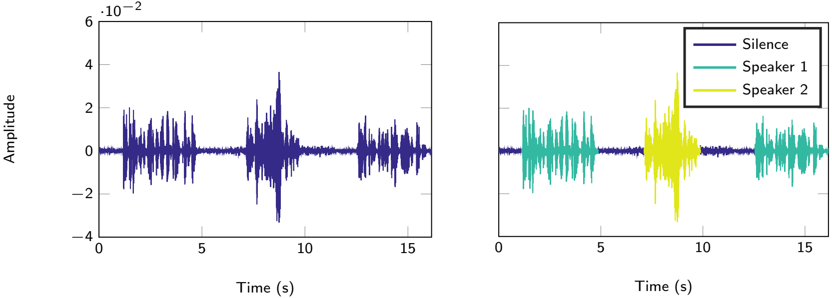
Bonn, Germany
LISTEN Project Workshop

July, 2018

[1] now with Cirrus Logic

# Introduction

- Speaker diarization $\Rightarrow$ segment audio into homogeneous sections with only one active speaker
  - "**who spoke when?**"



# Applications of diarization

- Annotation of meeting transcripts with speaker labels
  - Attorney meetings, corporate/business meetings
- Improve performance of Automatic Speech Recognition (ASR) systems by allowing effective speaker acoustic model adaptation

# Speaker diarization approaches

Distant speech diarization

- Can we use signal characteristics of the voice?
- Can we use the position of the sound source?
  - both are affected by noise, reverberation and non-speech

Two main approaches

- Single-microphone approaches are usually based on spectral differences
- Multi-microphone approaches include spatial information

Clustering

- either start with many clusters which are then merged successively until a stopping criteria is reached
- or start with only one cluster and split into new clusters until a stopping criteria is reached

# Multichannel approaches

Diarization can exploit spatial information in the multichannel case either by

- estimating TDOAs - time delay of the same signal at two different microphones, or
- estimating DOAs by maximizing e.g. steered response power

## TDOA Estimation

$$G_{PHAT}(f) = \frac{Y_1(f) \cdot Y_2^*(f)}{|Y_1(f) \cdot Y_2^*(f)|}$$

$Y_{1,2}(f)$ are the Fourier transforms of input signals.
TDOA at frame $l$ is found from

$$\tau_l = \operatorname*{argmax}_{\tau} R_{PHAT}(\tau)$$

where $R_{PHAT}(\tau)$ is the inverse Fourier transform of $G_{PHAT}(f)$

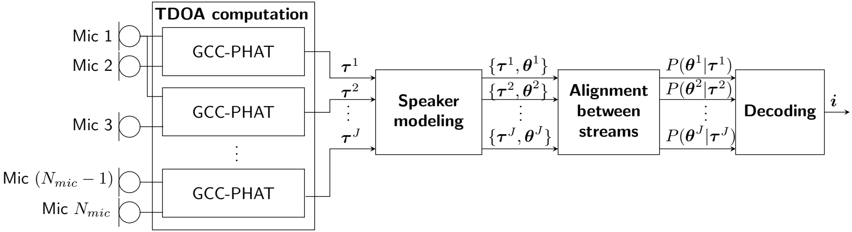
# Problem Addressed

TDOA estimation performance for distance speech diarization is degraded by

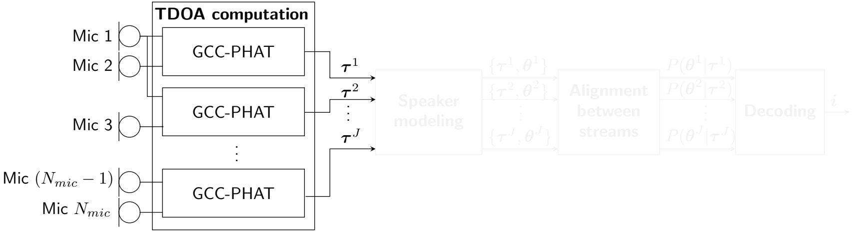
- reverberation
- noise
- VAD errors
- overlapping talkers
- non-speech (e.g. door closing)

Aim to build statistical models of the source TDOAs robust to erroneous data

# Proposed method

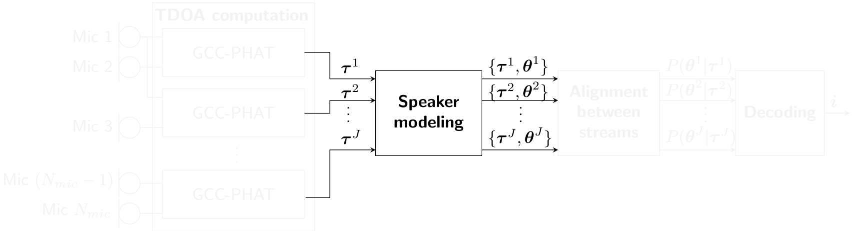


# TDOA computation




- $N_{mic}$ microphones: $J = 0.5 \cdot N_{mic} \cdot (N_{mic} - 1)$ TDOA streams
- The TDOA for frame $l$ and stream $j$ is denoted $\tau_l^j$
  - frames of 500 ms with 87.5% overlap
- A TDOA stream $\boldsymbol{\tau}^j$ is created by concatenating all per-frame TDOAs $\tau_l^j$

# Speaker modeling



# Speaker modeling

- Gaussian Mixture Model (GMM) for each mixture $i$, stream $j$

$$\boldsymbol{\theta}_i^j = (\lambda_i^j, \mu_i^j, \sigma_i^j)$$

- $N_{spk} + 1$ mixtures are considered
  - $N_{spk}$ mixtures to model the speakers' TDOAs
  - An additional mixture $\boldsymbol{\theta}_B^j$ to model the noisy estimates

**Problem**

The Expectation Maximization (EM) can be used to obtain $\boldsymbol{\theta}$, however in common applications, $\boldsymbol{\tau}^j$ can be inaccurate due to reverberation, noise, non-speech acoustic events

**Proposed solution**

Linear constraints on the mean and the standard deviation in the EM algorithm are included to estimate $\boldsymbol{\theta}$ robustly to these erroneous TDOA estimates

# Speaker modelling - Constraints on the mean

- Linear constraints on the distribution means:
  - The mean of the noise mixture, $\mu_B$, is **independent** of the speakers' means (defined with matrix $\boldsymbol{\mathcal{M}}$)
  - The speakers' means are separated by a **minimum distance** to avoid them being determined unreasonably close to each other (defined with vector $\boldsymbol{\mathcal{C}}$)

$$\boldsymbol{\mu} = \boldsymbol{\mathcal{M}}\boldsymbol{\beta} + \boldsymbol{C} \Rightarrow \begin{bmatrix} \mu_B \\ \mu_1 \\ \mu_2 \\ \dots \\ \mu_{N_{spk}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ C_2 \\ \dots \\ C_{N_{spk}} \end{bmatrix}$$

# Speaker modeling - Constraints on the standard deviation

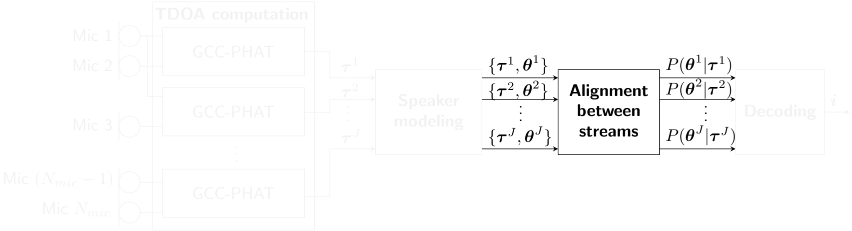
- Linear constraints on the standard deviation:
  - The variance of the noise mixture is **greater** than the variance of the speakers' mixtures (defined with matrix $\mathcal{G}$)
  - Variance of all speakers TDOAs (e.g. due to head movements) assumed to be similar $\Rightarrow$ the standard deviation of every speakers' mixture is the **same** (defined with matrix $\mathcal{G}$)

$$\boldsymbol{\iota} = \boldsymbol{\mathcal{G}\Upsilon} \Rightarrow \begin{bmatrix} 1/\sigma_B \\ 1/\sigma_1 \\ ... \\ 1/\sigma_{N_{spk}} \end{bmatrix} = \begin{bmatrix} \iota_B \\ \iota_1 \\ ... \\ \iota_{N_{spk}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ ... & ... \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \Upsilon_1 \\ \Upsilon_2 \end{bmatrix}$$

  - Additionally, variance upper and lower bounds (1.25 ms and 0.03125 ms respectively) are applied to avoid unlikely values
- Parameter estimation is performed using Expectation Constrained Maximization and Minorization-Maximization[2]

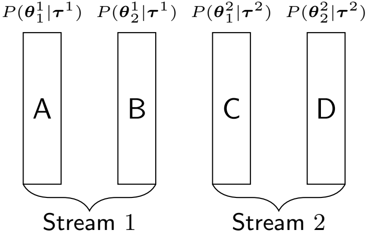
[2] Didier Chauveau, David Hunter. "ECM and MM algorithms for normal mixtures with constrained parameters", 2013. Available Online: https://hal.archives-ouvertes.fr/hal-00625285v2

# Alignment between streams



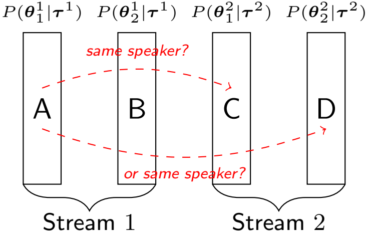
# Alignment between streams

- Alignment to ensure that the $N_{spk}$ speaker indexes represent the **same speaker** across the different $J$ streams for frames $l$.



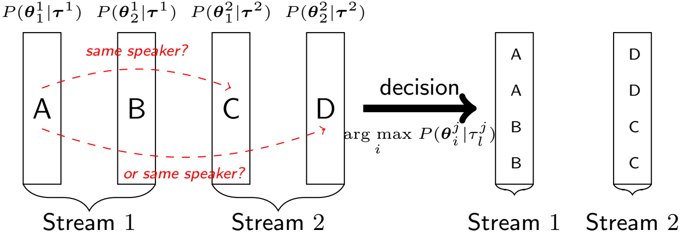
# Alignment between streams

- Alignment to ensure that the $N_{spk}$ speaker indexes represent the **same speaker** across the different $J$ streams for frames $l$.



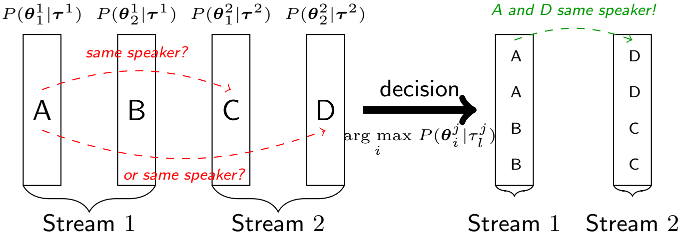
# Alignment between streams

- Alignment to ensure that the $N_{spk}$ speaker indexes represent the **same speaker** across the different $J$ streams for frames $l$.



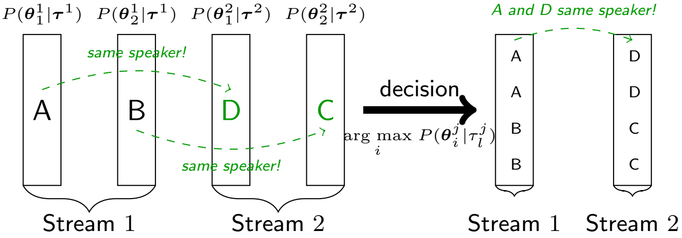
# Alignment between streams

- Alignment to ensure that the $N_{spk}$ speaker indexes represent the **same speaker** across the different $J$ streams for frames $l$.

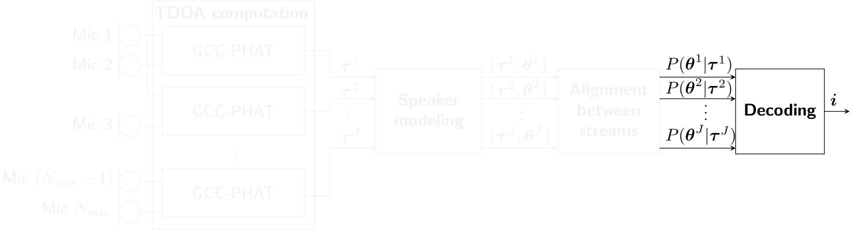


# Alignment between streams

- Alignment to ensure that the $N_{spk}$ speaker indexes represent the **same speaker** across the different $J$ streams for frames $l$.



# Decoding


TDOA computation
Mic 1
Mic 2
Mic 3
Mic $(N_{mic}-1)$
Mic $N_{mic}$
GCC-PHAT
GCC-PHAT
GCC-PHAT
$\tau^1$
$\tau^2$
$\tau^J$
Speaker modeling
$\{\tau^1, \theta^1\}$
$\{\tau^2, \theta^2\}$
$\{\tau^J, \theta^J\}$
Alignment between streams
$P(\boldsymbol{\theta}^1|\boldsymbol{\tau}^1)$
$P(\boldsymbol{\theta}^2|\boldsymbol{\tau}^2)$
$P(\boldsymbol{\theta}^J|\boldsymbol{\tau}^J)$
Decoding
$\boldsymbol{i}$

# Decoding

- The aim of the decoding block is to find, for each frame $l$, the speaker index $i$ that **maximizes the posterior probability** of the speaker model $\boldsymbol{\theta}_i^j$ given the TDOA sample $\tau_l^j$ as $\arg\max_i P(\boldsymbol{\theta}_i^j|\tau_l^j)$, where,

$$P(\boldsymbol{\theta}_i^j|\tau_l^j) = \frac{P(\tau_l^j|\boldsymbol{\theta}_i^j) \cdot P(\boldsymbol{\theta}_i^j)}{\sum_{e=1}^{N_{spk}} P(\tau_l^j|\boldsymbol{\theta}_e^j) \cdot P(\boldsymbol{\theta}_e^j)}$$

# Decoding - Approaches

1. **Stream selection** approach selects the optimal TDOA stream to employ for decoding based on the Bayesian Information Criterion (BIC) by maximizing

$$\mathrm{BIC}(\boldsymbol{\theta}^j, \boldsymbol{\tau}^j) = -2 \log \mathcal{L}(\boldsymbol{\theta}^j | \boldsymbol{\tau}^j) + N_{fp} \cdot \log(N_{TDOA})$$

   where:
   - $\mathcal{L}(\boldsymbol{\theta}^j | \boldsymbol{\tau}^j)$ is the likelihood of the model $\boldsymbol{\theta}^j$ given the data $\boldsymbol{\tau}^j$
   - $N_{fp}$ is the number of free parameters to be estimated in $\boldsymbol{\theta}$

2. **Stream combination** approach computes the average of the probabilities over all $J$ streams and selects $i$ as

$$\operatorname*{argmax}_i \frac{1}{J} \sum_{j=1}^{J} P(\boldsymbol{\theta}_i^j | \tau_l^j), \text{ where } i = \{1, \cdots, N_{spk}\}$$

# Decoding - HMM

- A **Hidden Markov Model** (HMM) is introduced to avoid very unlikely short utterances
- Each state of the HMM represents one speaker and all the states are interconnected
- **Transition probabilities** for speakers $q$ and $r$ are chosen as

$$1/(1-a_{qq}) = \text{ average duration in frames of speaker } q$$
$$a_{qr} = (1-a_{qq})/N_{spk}-1)$$

- **Observation probabilities** are set to $P(\boldsymbol{\theta}_i|\tau_l)$ for speaker $i$ at frame $l$
- Viterbi algorithm is applied to extract the speaker label estimate at frame $l$

# Evaluation

- Evaluated on distant multi-microphone partition of **NIST RT-05**
- The baseline used to compare the performance is **DiarTK**[3]
  - Open source toolkit where the clusters are merged depending on a mutual information loss
  - It was given TDOA streams from all microphone pairs $\boldsymbol{\tau}$
- In both systems $\boldsymbol{N_{spk}}$ is set to 10
- The scoring is restricted to **speech active regions**
  - The relative reduction of the speaker error (RRSE) time is used

$$\text{RRSE} = \frac{SE_{baseline} - SE_{proposed}}{SE_{baseline}} \cdot 100(\%)$$

[3] D. Vijayasenan, F. Valente, and H. Bourlard (2011). "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization". In: *IEEE Trans. Audio, Speech, Lang. Process.* 19.2, pp. 431–438.

# Results

- Summary of the performance in terms of RRSE for the two proposed approaches

| Meeting | $N_{spk}$ | $N_{mic}$ | Stream Selection | Stream Combination |
|---|---|---|---|---|
| AMI1 | 4 | 8 | 54.1 | 85.6 |
| AMI2 | 4 | 8 | -6.0 | 31.3 |
| CMU1 | 4 | 3 | 75.2 | 77.1 |
| CMU2 | 4 | 3 | 77.4 | 38.0 |
| ICSI1 | 7 | 6 | 84.6 | 70.8 |
| ICSI2 | 9 | 6 | 50.1 | 49.9 |
| NIST1 | 10 | 7 | -54.3 | -56.9 |
| NIST2 | 4 | 7 | 0.0 | 31.2 |
| VT1 | 5 | 2 | 8.3 | 8.3 |
| VT2 | 5 | 2 | 25.9 | 25.9 |
| Mean RRSE(%) | | | **31.5** | **36.1** |

# Conclusions

- A speaker diarization method was presented that uses:
  - Spatial features in the form of **TDOAs**
  - Features modelled to include **linear constraints** to increase robustness
- The evaluation of the proposed method was carried out on a distant multi-microphone database achieving **36.1% RRSE** with respect to DiarTK
- Further improvements can be gained when the **number of speakers is known** *a priori* (RRSE of 51.9%)