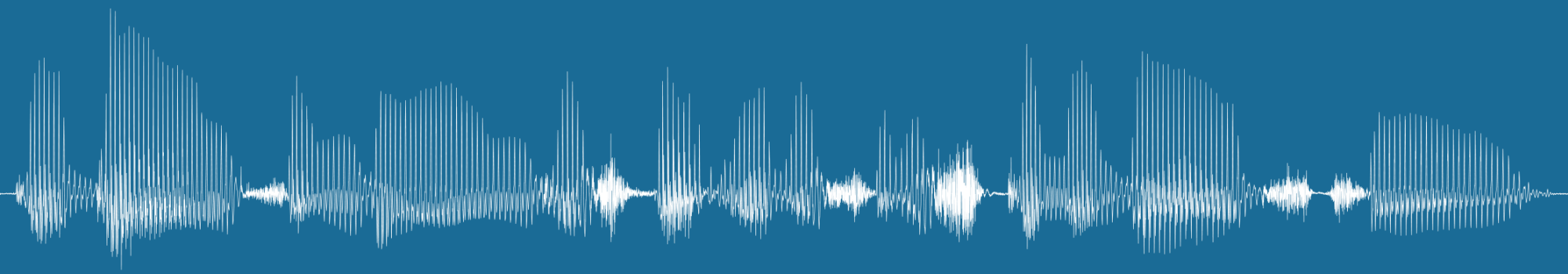
Imperial College London

UNIVERSITY OF Southampton



MULTICHANNEL OVERLAPPING SPEAKER SEGMENTATION USING MULTIPLE HYPOTHESIS TRACKING OF ACOUSTIC AND SPATIAL FEATURES

June, 2021

Aidan Hogg, Christine Evers and Patrick Naylor

Electrical and Electronic Engineering, Imperial College London, UK

Electronics and Computer Science, University of Southampton, UK

Imperial College London

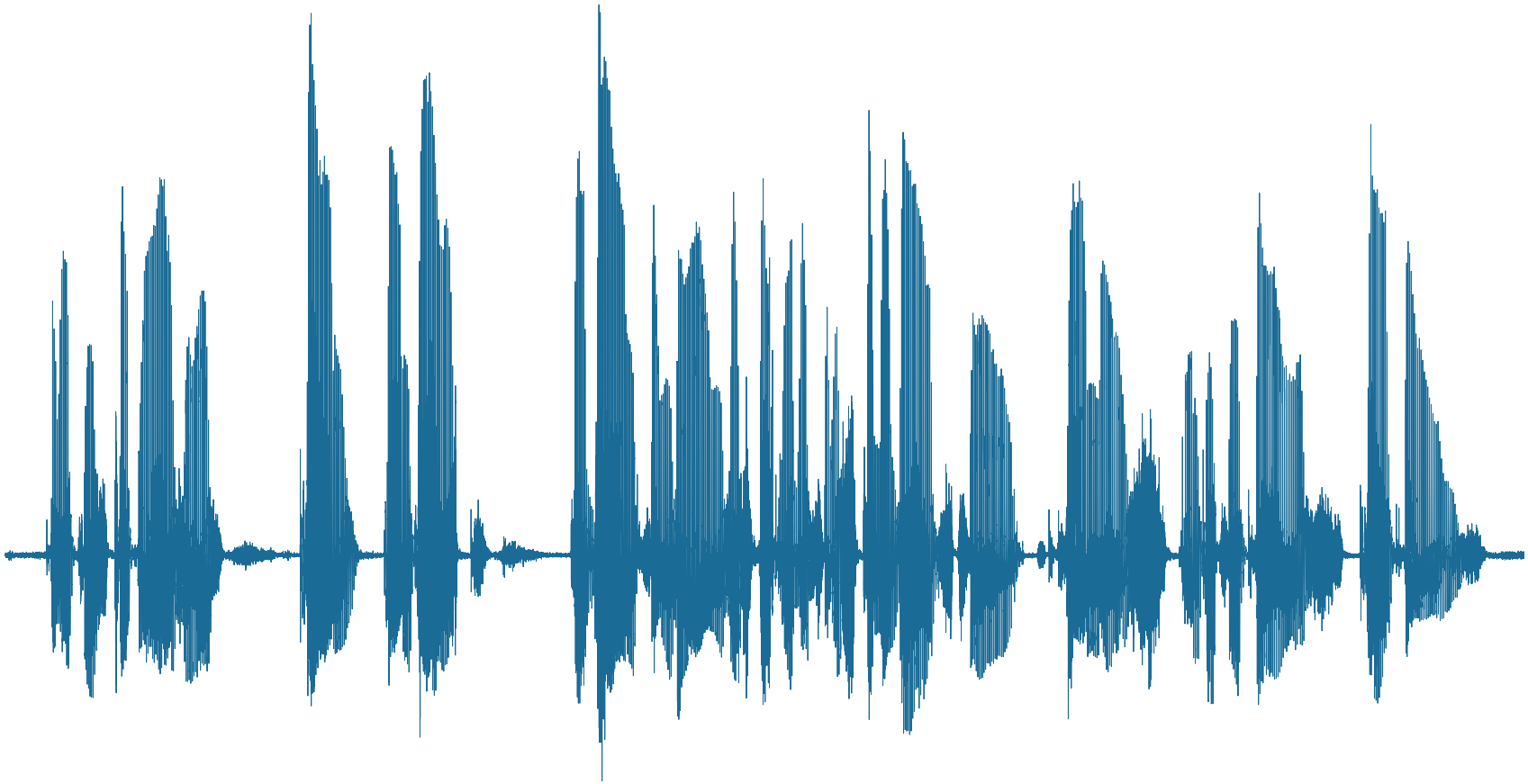
What is speaker diarization?

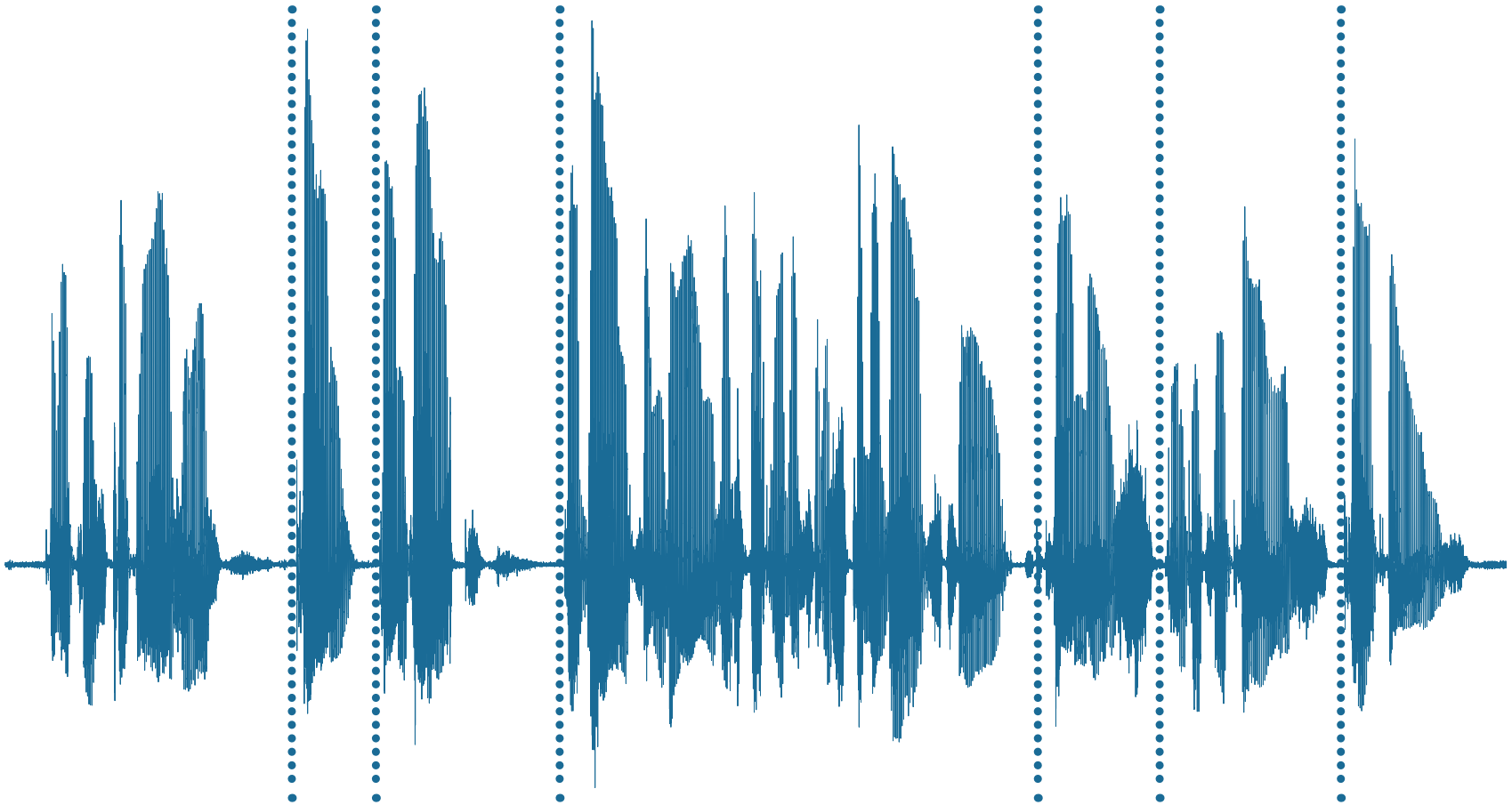
Answers the question “who spoke when?” in an audio recording.

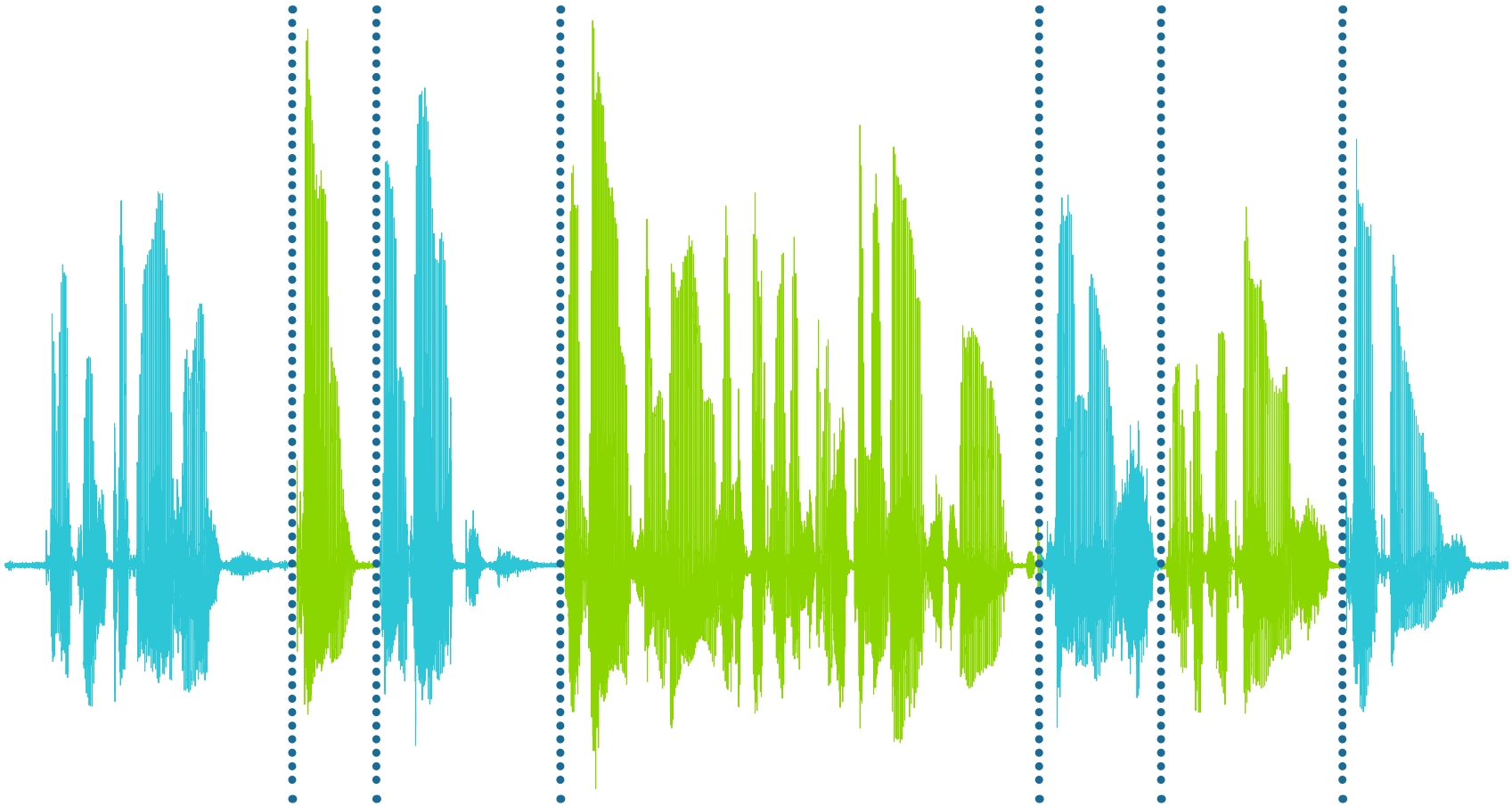
Is diarization really that useful?

- Speaker indexing and rich transcription
- Speaker segmentation and clustering helping Automatic Speech Recognition (ASR) systems
- Preprocessing modules for single speaker-based algorithms

DIARIZATION METHOD

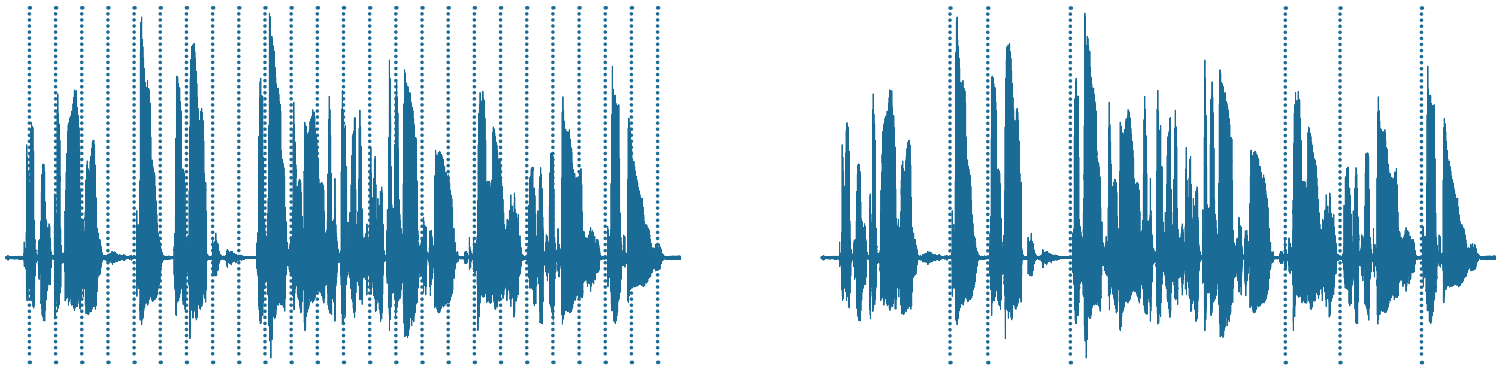






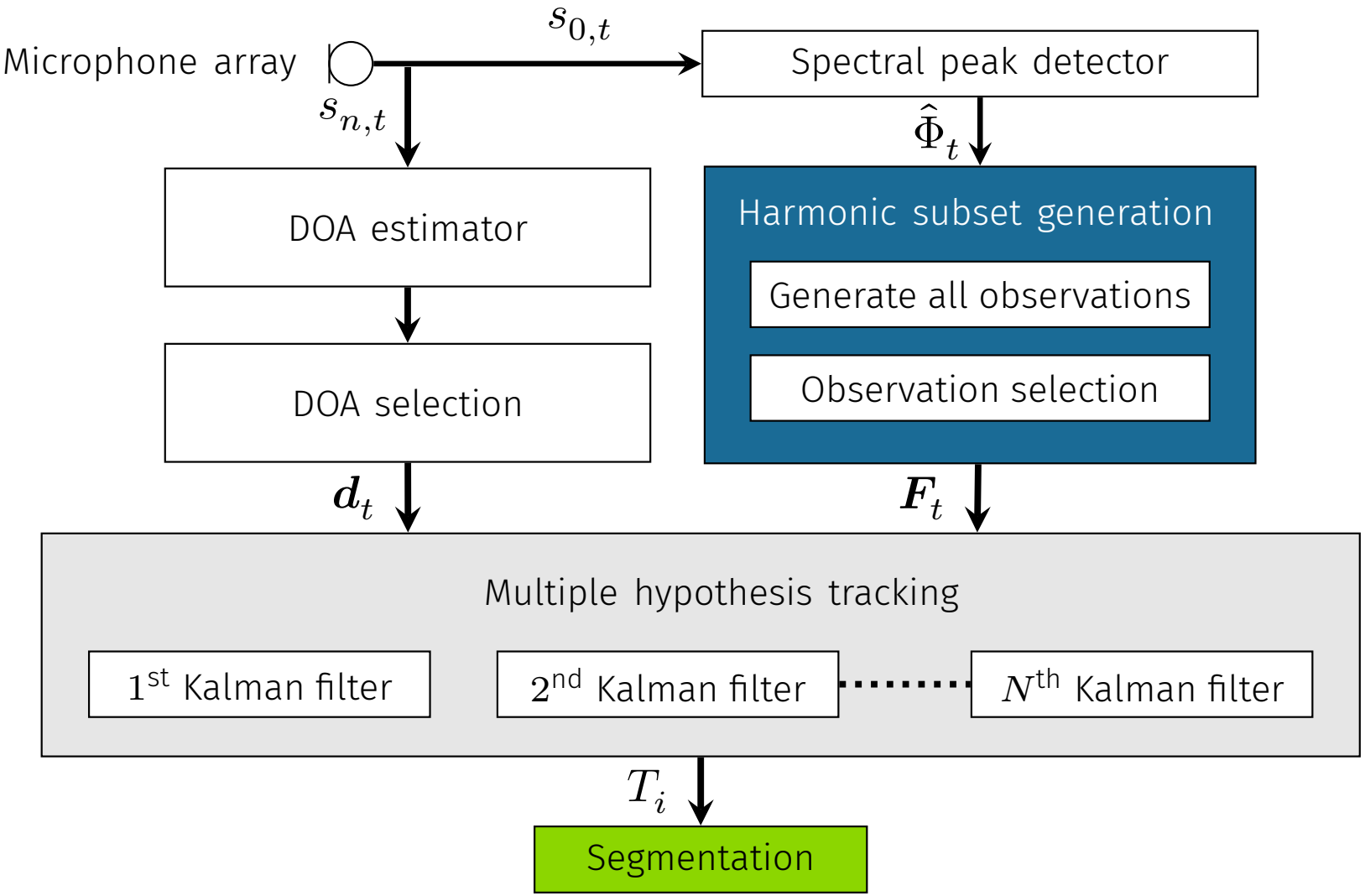
Is good segmentation really that useful?

Why not just segment the audio stream into small uniform segments and cluster with realignment?



If the speech segments are small then each segment only contains a small amount of information that can be used for clustering.

PROPOSED SYSTEM ARCHITECTURE


Microphone array
$s_{0,t}$
$s_{n,t}$
Spectral peak detector
$\hat{\Phi}_t$
DOA estimator
DOA selection
d_t
Harmonic subset generation
Generate all observations
Observation selection
F_t
Multiple hypothesis tracking
1st Kalman filter
2nd Kalman filter
N^{th} Kalman filter
T_i
Segmentation

In this work all the harmonics of voiced speech are tracked along with the DOA estimates so that overlapping speech can be considered.

F_0 harmonics and DOA observation:

$$\boldsymbol{z}_{t,n} = [\boldsymbol{f}_{t,m}, d_{t,v}]^T,$$

The state equation for the i^{th} speaker:

$$\boldsymbol{x}_{i,t} = \boldsymbol{x}_{i,t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_t),$$

where

$$\boldsymbol{x}_i = \left[x_f, x_{\text{d}}\right]^T, \quad \boldsymbol{Q}_t = \text{diag}(\sigma_w^2, \sigma_w^2).$$

Observation:

$$\boldsymbol{z}_{t,n} = \boldsymbol{H}_{t,n}\boldsymbol{x}_{i,t} + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim \mathcal{N}(\mathbf{o}, \boldsymbol{R}_t),$$

where

$$\boldsymbol{H}_{t,n} = \begin{bmatrix} h_{t,n}(0) & h_{t,n}(1) & \cdots & h_{t,n}(K) & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T,$$

and

$$\boldsymbol{R}_t = \text{diag}(\sigma_v^2, \cdots, \sigma_v^2).$$

Prediction step:

$$\hat{\boldsymbol{x}}_{i,t|t-1} = \hat{\boldsymbol{x}}_{i,t-1|t-1},$$

$$\boldsymbol{P}_{i,t|t-1} = \boldsymbol{P}_{i,t-1|t-1} + \boldsymbol{Q}_t.$$

Update step: (performed if an observation exists)

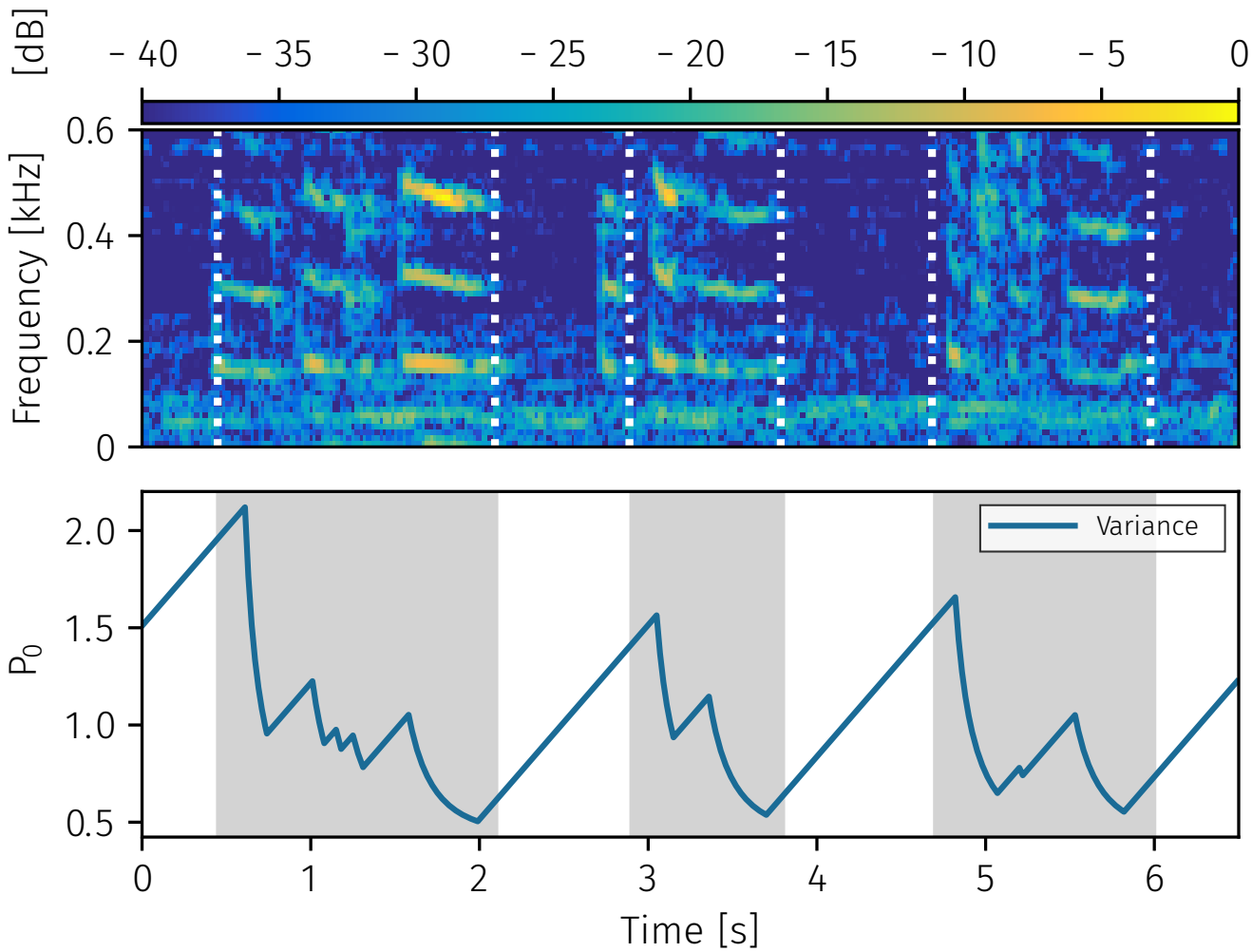
$$\hat{\boldsymbol{x}}_{i,t|t} = \hat{\boldsymbol{x}}_{i,t|t-1} + \boldsymbol{k}_{i,t}(\boldsymbol{z}_{t,n} - \boldsymbol{H}_{t,n}\hat{\boldsymbol{x}}_{i,t|t-1}),$$
$$\boldsymbol{P}_{i,t|t} = (\boldsymbol{I} - \boldsymbol{k}_{i,t}\boldsymbol{H}_{t,n})^2 \boldsymbol{P}_{i,t|t-1} + \boldsymbol{k}_{i,t}\boldsymbol{R}_t\boldsymbol{k}_{i,t}^T.$$

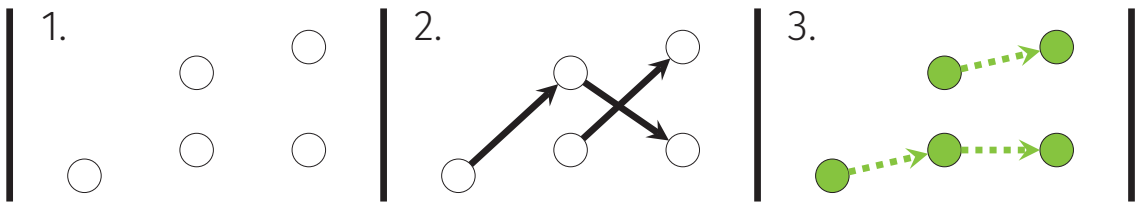
If $\boldsymbol{k}_{i,t} = [1, \cdots, 1] \implies \hat{\boldsymbol{x}}_{i,t|t} = \boldsymbol{z}_{t,n}$ (just the measurement)

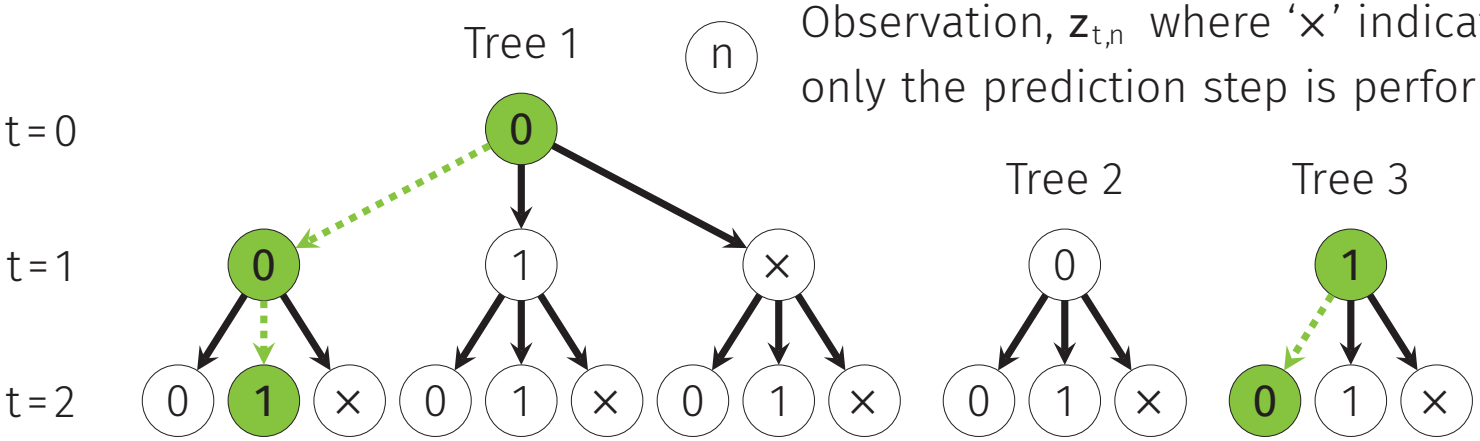
If $\boldsymbol{k}_{i,t} = [0, \cdots, 0] \implies \hat{\boldsymbol{x}}_{i,t|t} = \hat{\boldsymbol{x}}_{i,t|t-1} - 1$ (just the prediction)

Optimal Kalman gain:

$$\boldsymbol{k}_{i,t} = \frac{\boldsymbol{P}_{i,t|t-1}\boldsymbol{H}_{t,n}^T}{\boldsymbol{S}_{i,t}}$$

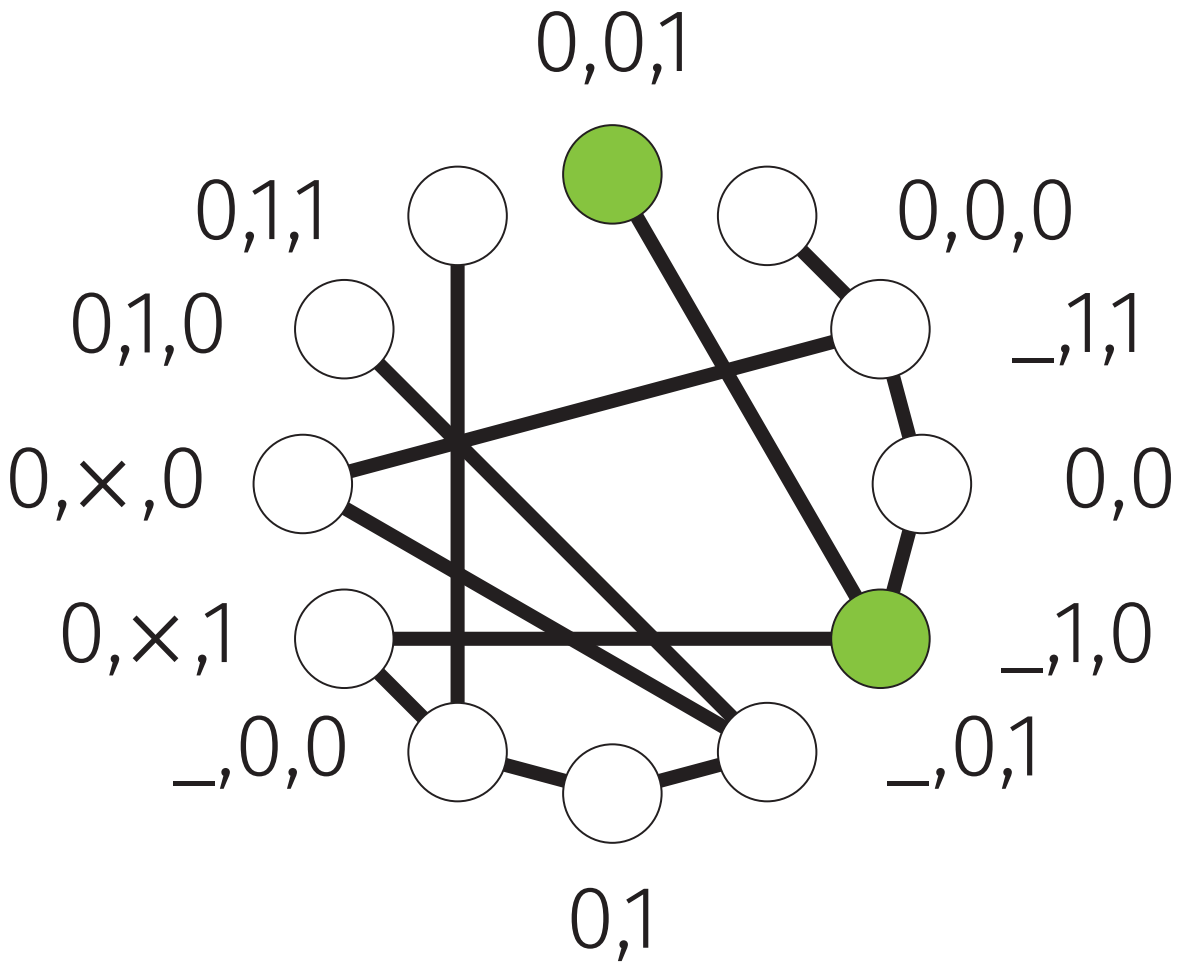

[dB]
− 40
− 35
− 30
− 25
− 20
− 15
− 10
− 5
0
Frequency [kHz]
0.6
0.4
0.2
0
P_0
2.0
1.5
1.0
0.5
Variance
0
1
2
3
4
5
6
Time [s]





Observation, $z_{t,n}$ where '×' indicates that only the prediction step is performed.

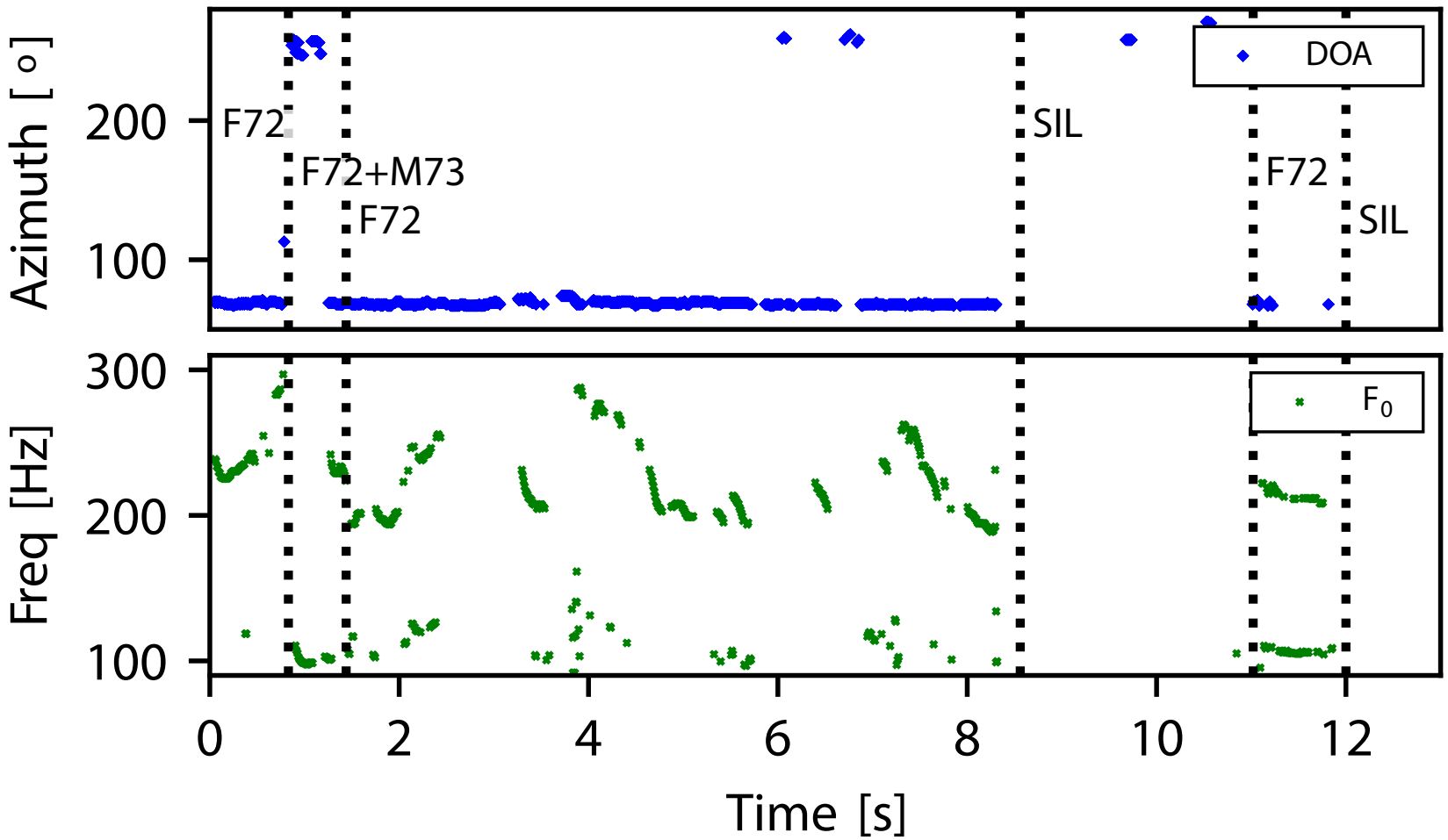
Imperial College London

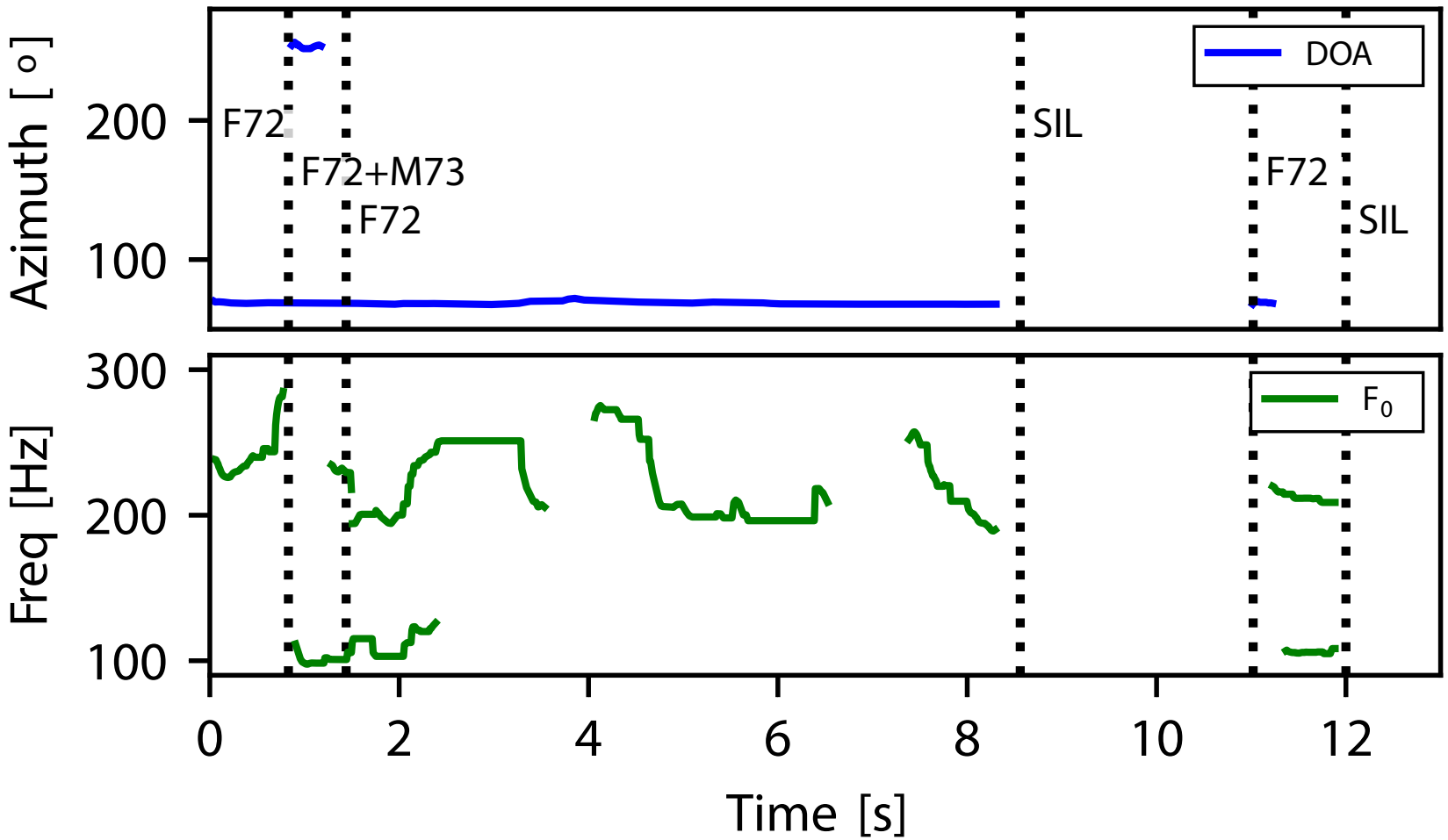


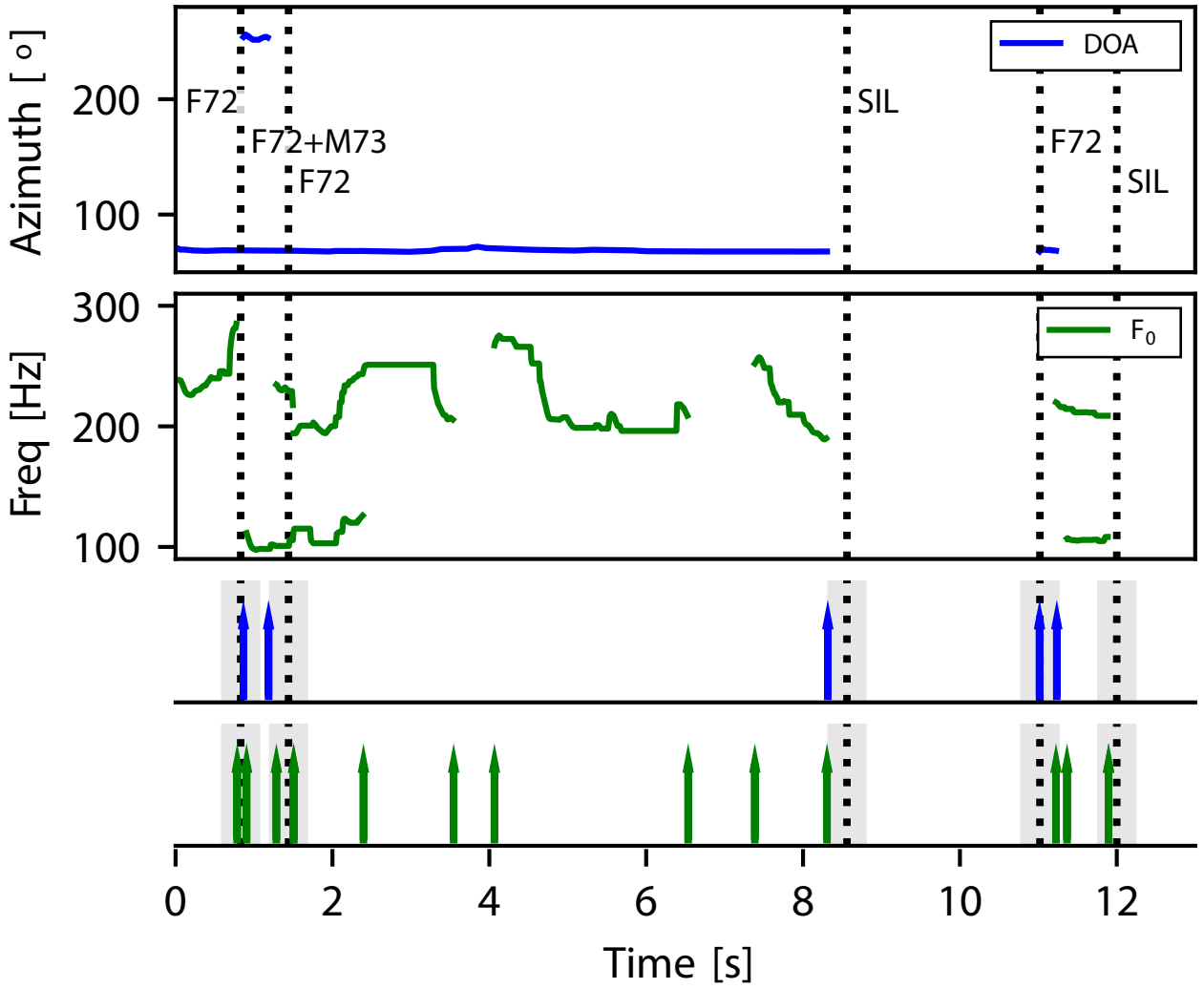
ILLUSTRATIVE EXAMPLE

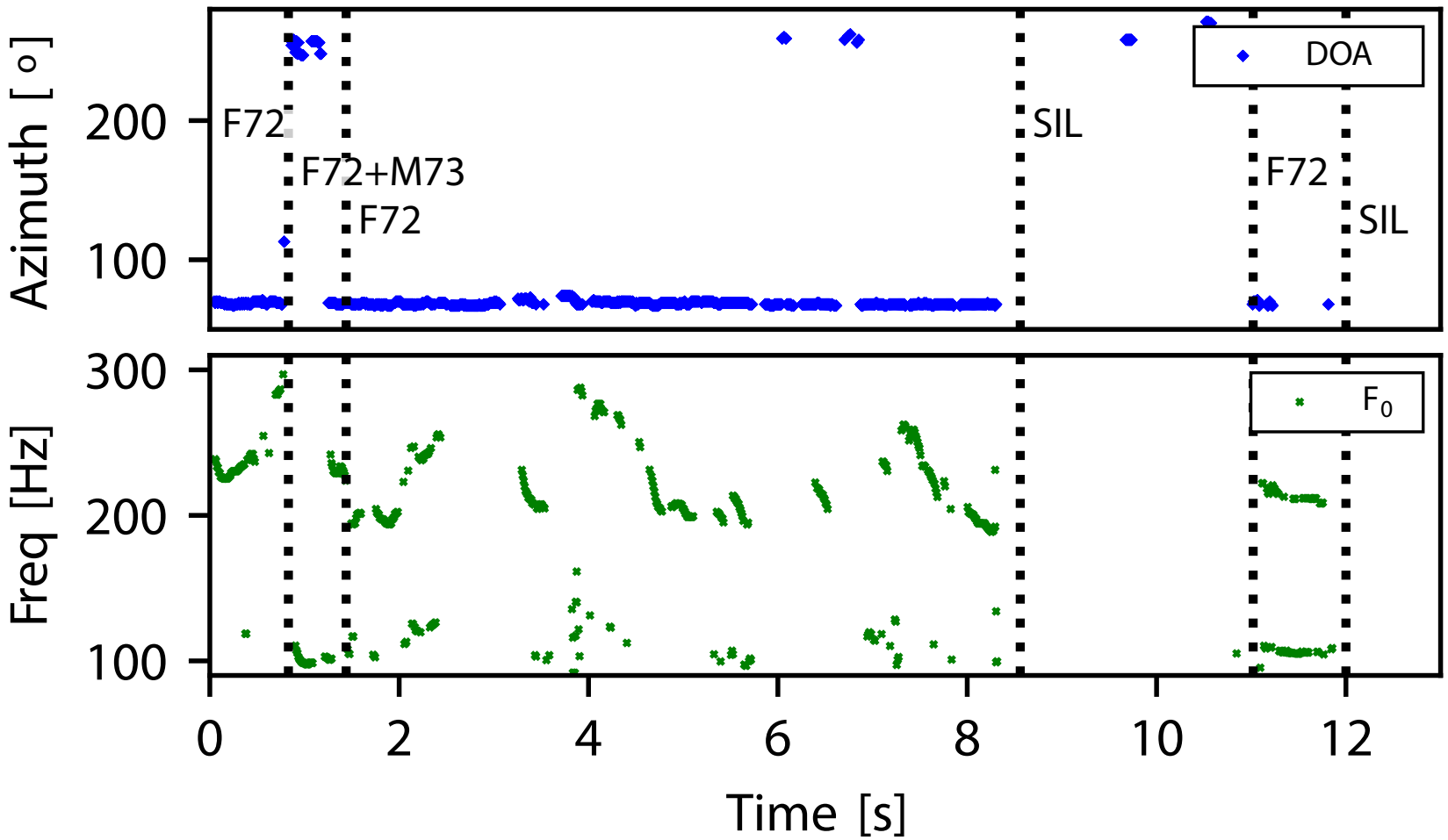
Multi-modal data set consisting of 100 hours of meeting recordings.

Recorded in English using three different rooms with different acoustic properties and includes mostly non-native speakers.


Azimuth [°]
200
100
F72
F72+M73
F72
SIL
F72
SIL
DOA
Freq [Hz]
300
200
100
F_0
0
2
4
6
8
10
12
Time [s]


Azimuth [°]
200
100
F72
F72+M73
F72
SIL
F72
SIL
DOA
Freq [Hz]
300
200
100
F_0
0
2
4
6
8
10
12
Time [s]

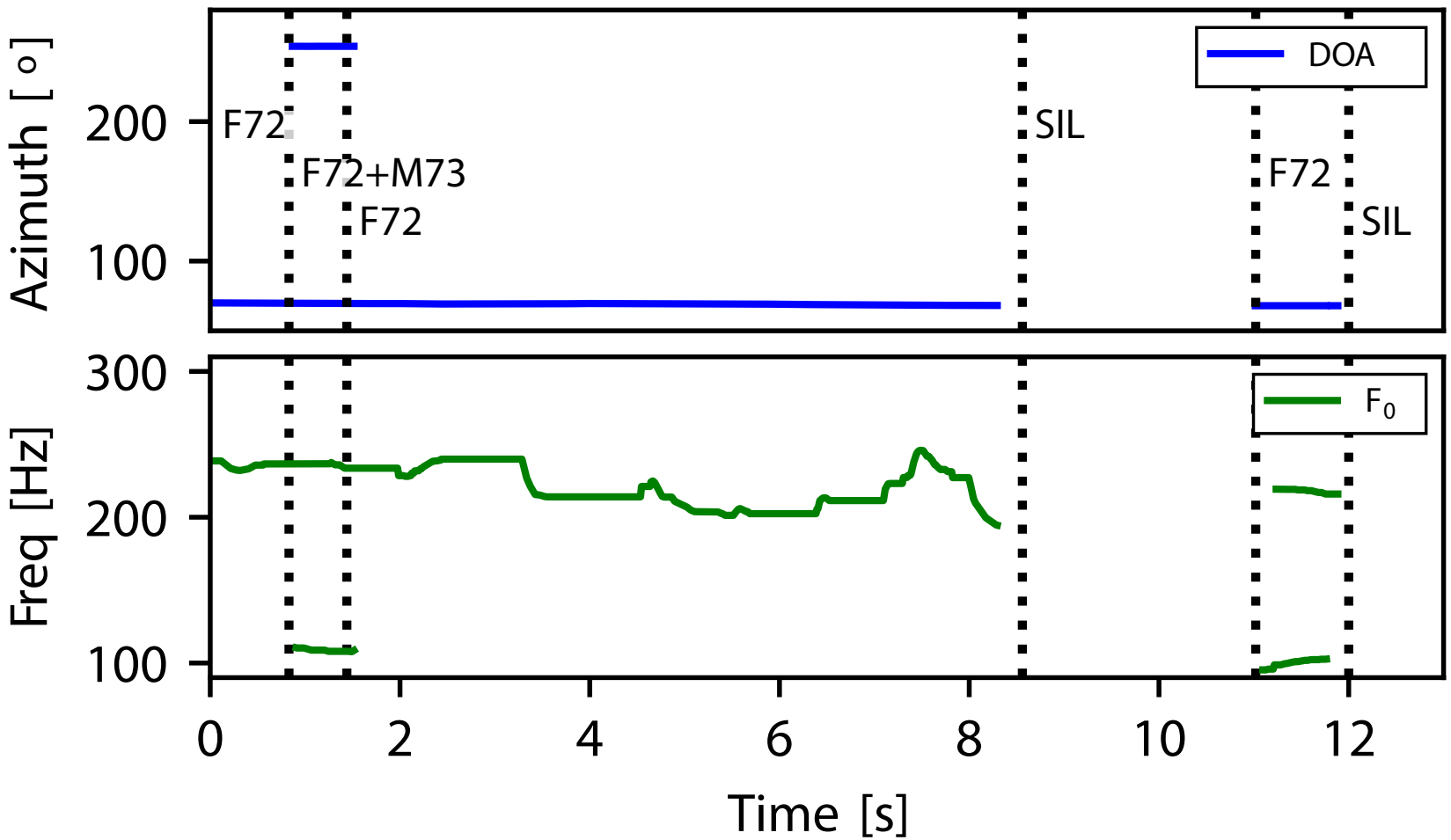

Azimuth [°]
200
100
DOA
F72
F72+M73
F72
SIL
F72
SIL
Freq [Hz]
300
200
100
F_0
0
2
4
6
8
10
12
Time [s]

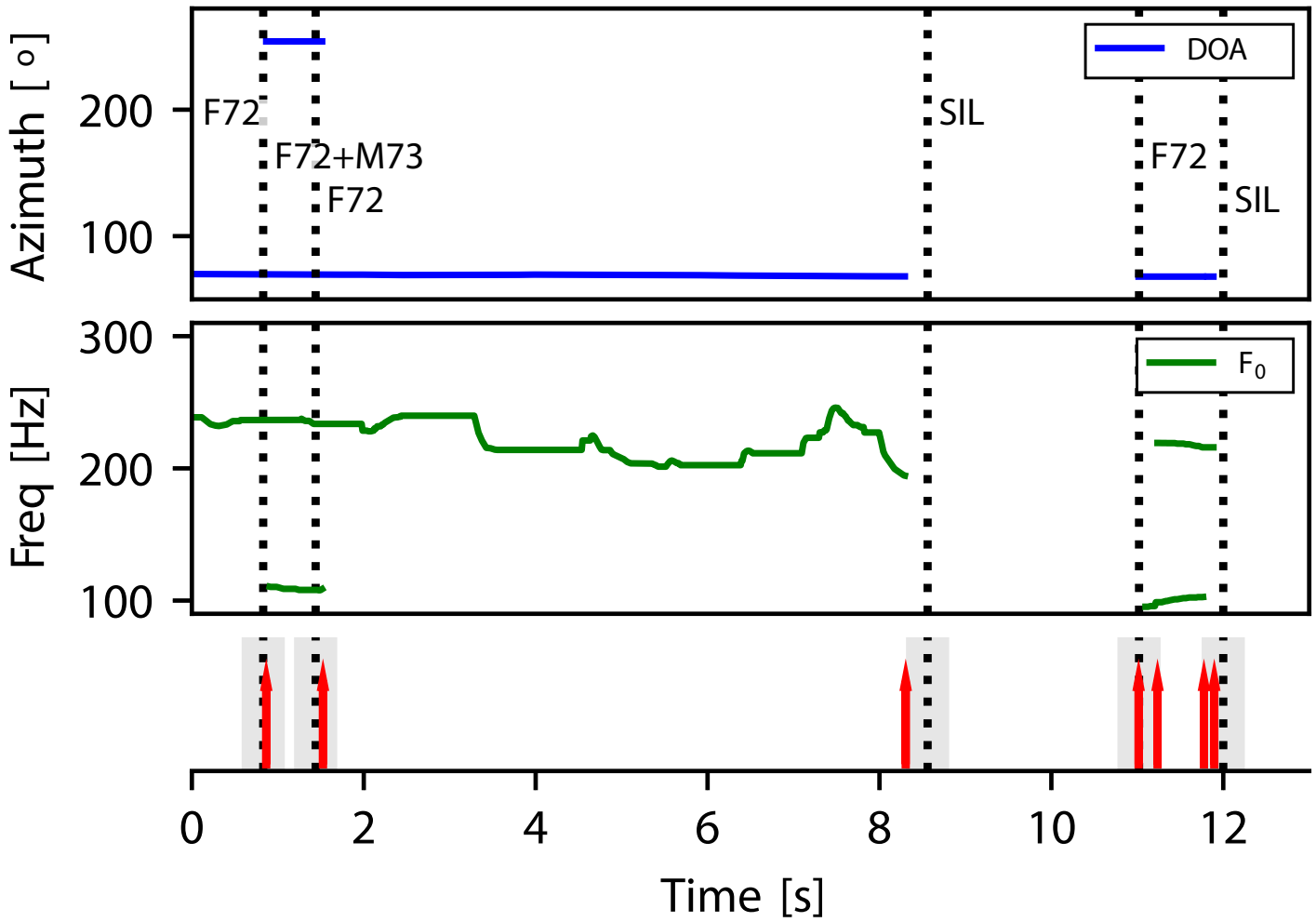

Azimuth [°]
200
100
F72
F72+M73
F72
SIL
F72
SIL
DOA
Freq [Hz]
300
200
100
F_0
0
2
4
6
8
10
12
Time [s]

DOA & PITCH ESTIMATES TRACKED SIMULTANEOUSLY

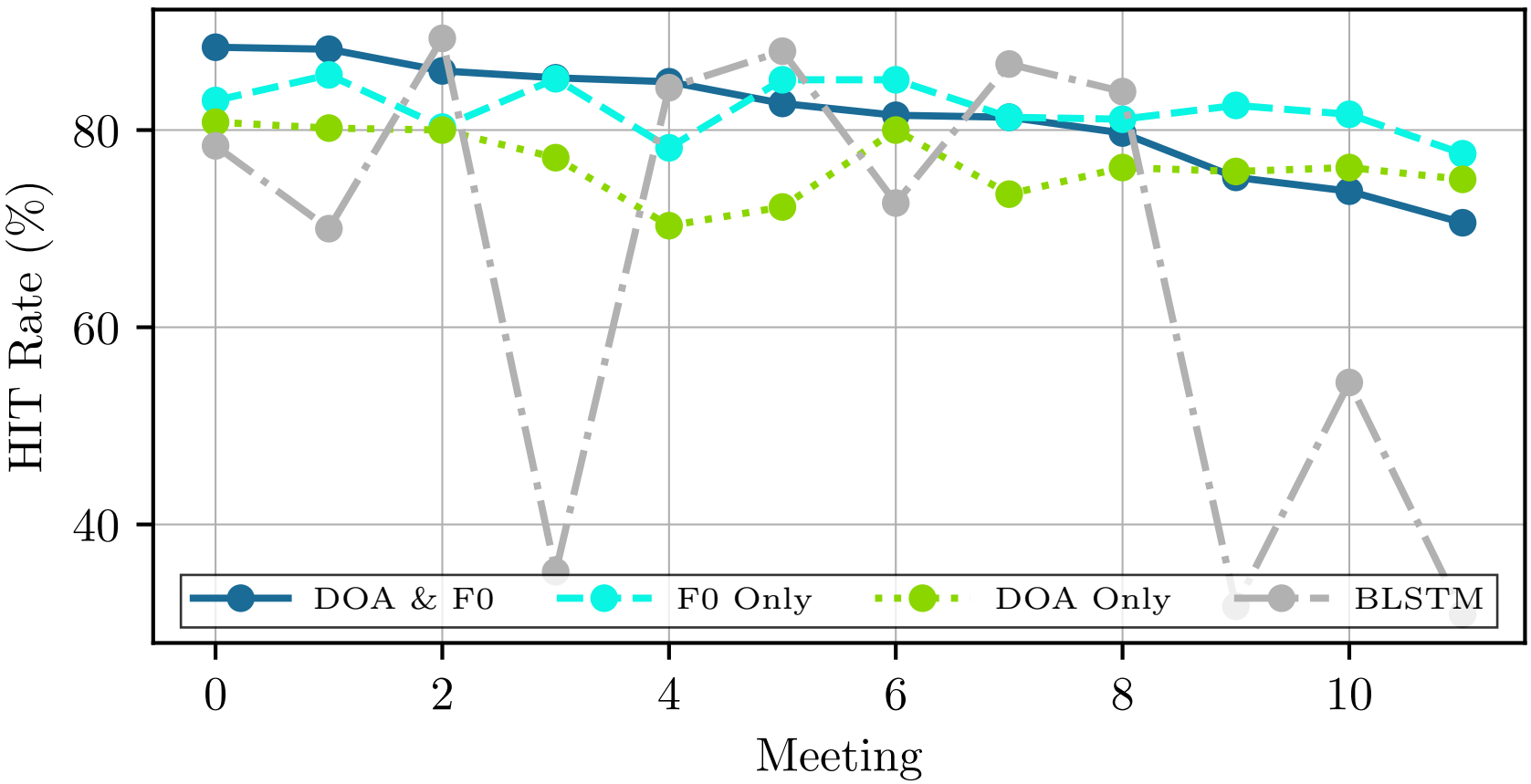
Illustrative Example

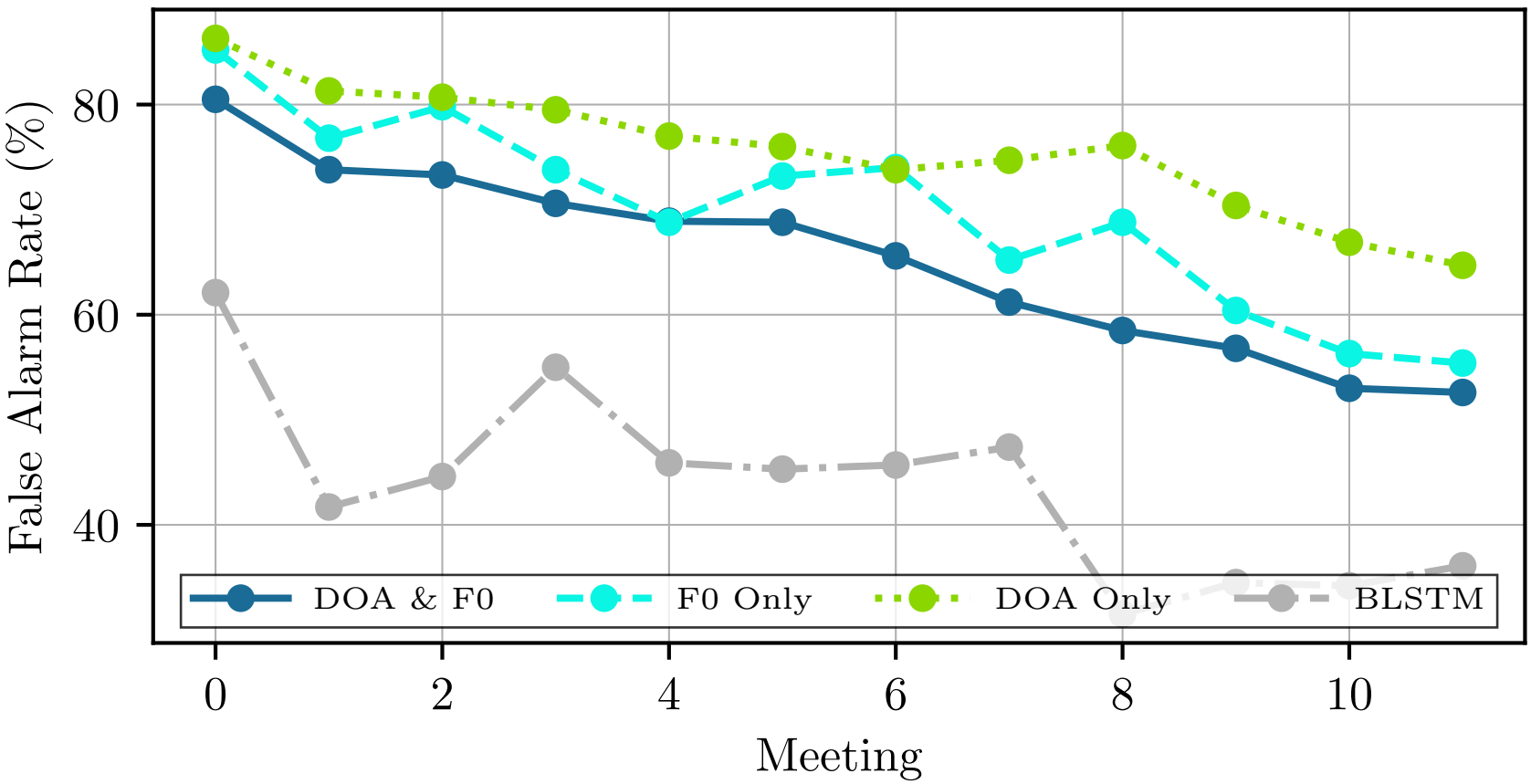
Imperial College London

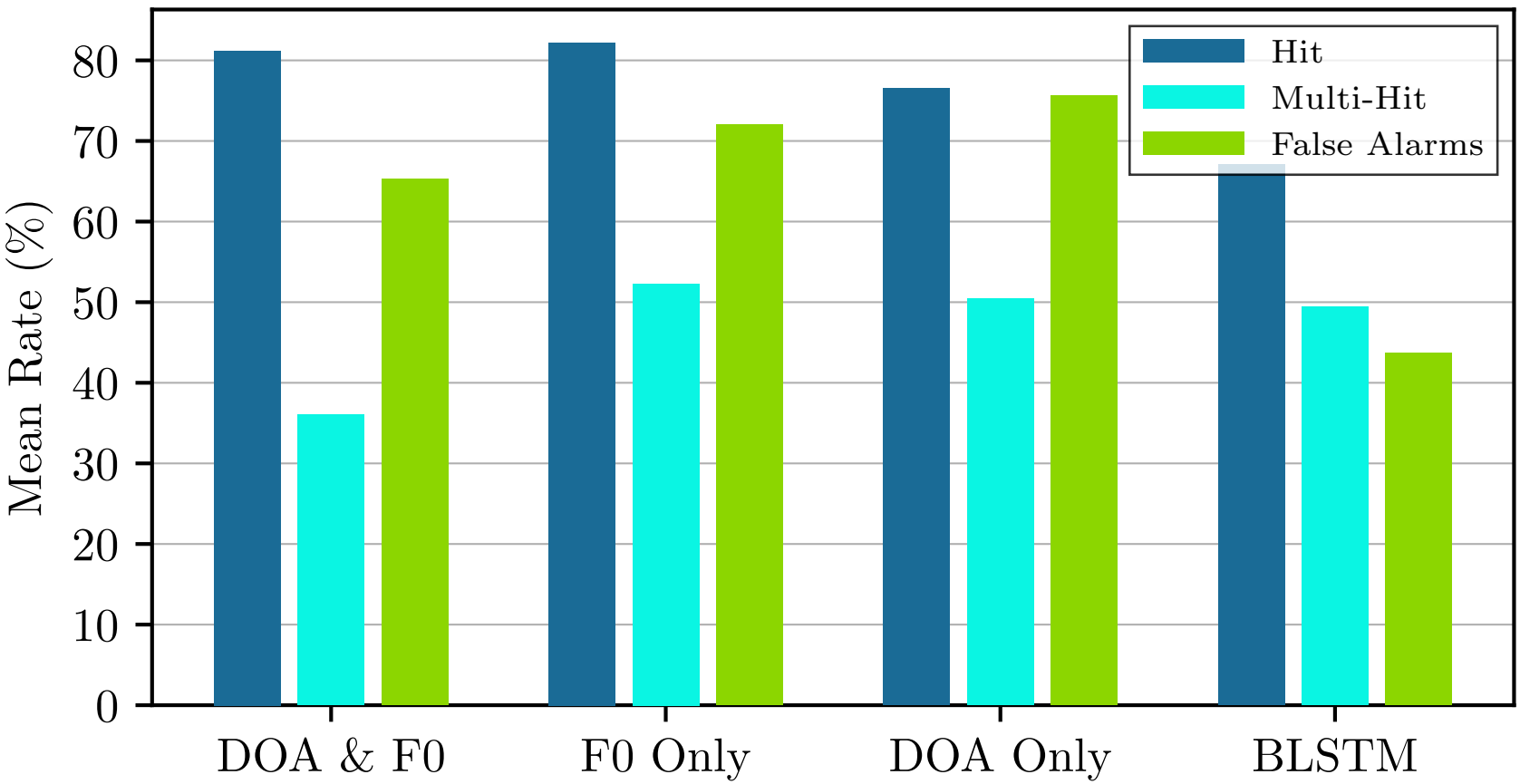



Azimuth [°]
200
100
F72
F72+M73
F72
SIL
F72
SIL
DOA
Freq [Hz]
300
200
100
F_0
0
2
4
6
8
10
12
Time [s]

RESULTS


HIT Rate (%)
80
60
40
0
2
4
6
8
10
Meeting
DOA & F0
F0 Only
DOA Only
BLSTM


False Alarm Rate (%)
80
60
40
0
2
4
6
8
10
Meeting
DOA & F0
F0 Only
DOA Only
BLSTM


Mean Rate (%)
0
10
20
30
40
50
60
70
80
DOA & F0
F0 Only
DOA Only
BLSTM
Hit
Multi-Hit
False Alarms

CONCLUSION

In this paper we have...

...proposed a novel method that uses a MHT framework to track the F0 and DOA of multiple speakers simultaneously.

...shown that MHT of both the DOA and F0 can lead to an improved speaker segmentation performance over tracking just one of these features alone.

Questions?

Please email: aidan.hogg13@imperial.ac.uk

Imperial College London

University of Southampton

EPSRC
Engineering and Physical Sciences Research Council