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

Aidan O. T. Hogg, Christine Evers and Patrick A. Naylor

Imperial College London

University of Southampton

1. Overview & Proposed system

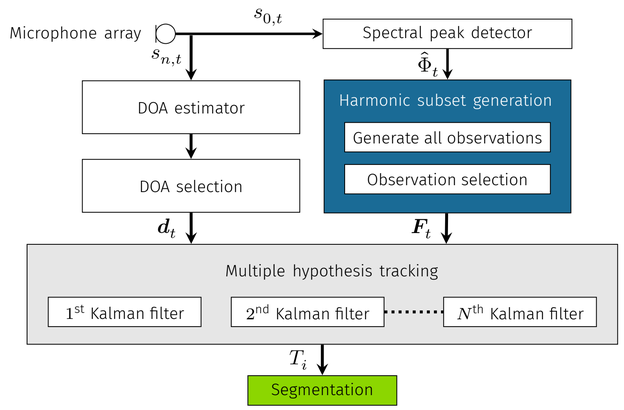
- **What does this paper propose?**

  A new multimodal approach for overlapping speaker segmentation that tracks simultaenously both the fundamental frequencies (F_0s) and the direction of arrivals (DOAs) of multiple, simultaneously active speakers.

- **How well does this proposed method perform?**

  The proposed multimodal method shows an improvement in segmentation performance compared to tracking features separately.

  Segmentation performance comparable to a deep learning approach is achieved but without the need for labelled training data and handles overlapping talkers.



2. Motivation

- **What is speaker diarization?**

  Answers the question "who spoke when?" and is required for applications such as, speaker indexing and automatic speech recognition (ASR).

- **Is performing segmentation before clustering beneficial?**

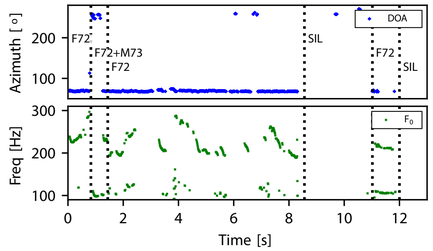
  If segmentation is performed before clustering then each segment will contain the maximum amount of information possible on a speaker's identity.

- **Why use direction of arrival for segmentation?**

  It has been shown in the past that spatial features can help improve speaker segmentation even in the context of overlapping speech.

- **Why use pitch for segmentation?**

  A previous study of meetings in the AMI corpus has shown that abrupt variations in voice pitch estimates are indicative of speaker changes.

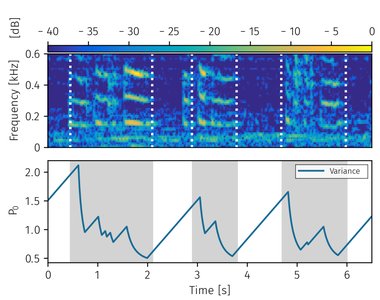


3. Kalman Filter For DOA & F_0 Tracking

All the harmonics of voiced speech are tracked along with the DOA estimates so that overlapping speech can be processed.

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

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

with observation:

$$\boldsymbol{z}_{t,n} = \boldsymbol{H}_{t,n}\boldsymbol{x}_{i,t} + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{o}, \boldsymbol{R}_t), \quad \text{where} \quad \boldsymbol{H}_{t,n} = \begin{bmatrix} h_{t,n}(0) & h_{t,n}(1) & \cdots & h_{t,n}(K) & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T,$$

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

$\boldsymbol{R}_t$ is the covariance of the observation noise.

Prediction step:

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

$\boldsymbol{Q}_t$ is the covariance of the process noise.

Update step:

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

Optimal Kalman gain:

$$\boldsymbol{k}_{i,t} = \boldsymbol{P}_{i,t|t-1}\boldsymbol{H}_{t,n}^T\boldsymbol{S}_{i,t}^{-1}, \quad \text{where} \quad \boldsymbol{S}_{i,t} = \boldsymbol{H}_{t,n}\boldsymbol{P}_{i,t|t-1}\boldsymbol{H}_{t,n}^T + \boldsymbol{R}_t.$$

Estimation error:

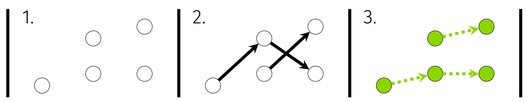
$$\boldsymbol{e}_{i,t|t} = \boldsymbol{z}_{t,n} - \boldsymbol{H}_{t,n}\hat{\boldsymbol{x}}_{i,t|t}.$$

4. Multiple Hypnosis Tracking

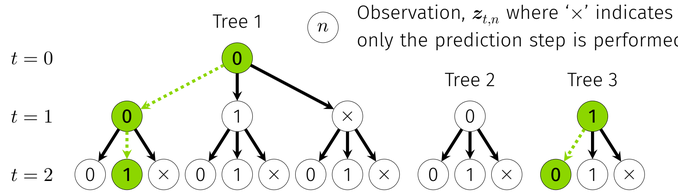
A MHT framework is exploited to simultaneously track both the F_0 and DOA features.

Tracking observations

1. Observations in time
2. Possible track
3. Best track



Multiple hypothesis tracking

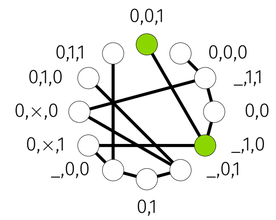


Observation, $\boldsymbol{z}_{t,n}$ where '×' indicates that only the prediction step is performed.

Maximum weighted clique (MWC)

Used to find the most likely set of tracks which do not conflict.

Each node is a track hypothesis and each edge connects 2 tracks which do not conflict. A score is assigned which is calculated by taking the average value of all previous estimation errors.



Track generation

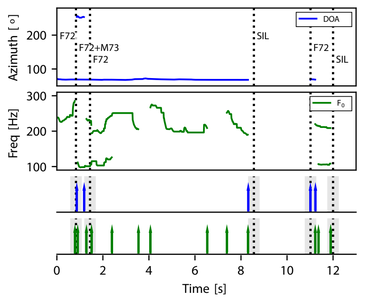
At each time-frame three possible tracks can be generated: a track only containing the F_0 observation; a track only containing the DOA observation and a track that fuses information from the F_0 and DOA observations.

The prediction step is executed for every time-frame. However, the update step is only performed when new observations emerge.

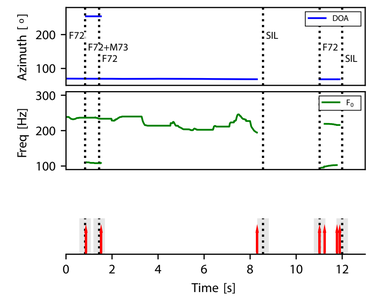
5. Results & Conclusion

Illustrative (AMI) example

Part of a meeting from the AMI corpus that contains overlapping speech.



Proposed method tracking DOA and F_0 tracked alone.



Proposed method tracking DOA and F_0 simultaneously.

Evaluation on AMI corpus

Performance comparison of the proposed method on the AMI corpus compared against the performance achieved by only using DOA or F0 features alone as well as a bidirectional long short term memory networks (BLSTM) approach [2].

Method	Hit	Miss	Multi-Hit	FA
Proposed	81.2%	18.8%	36.0%	65.3%
F_0 Only	82.2%	17.8%	52.3%	72.0%
DOA Only	76.5%	23.6%	50.4%	75.6%
BLSTM	67.1%	32.9%	49.4%	43.7%

Mean values across 12 meetings from the AMI corpus.

[2] H. Bredin *et al*, "pyannote.audio: neural building blocks for speaker diarization," in *IEEE Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP), 2020.*

Conclusion

- A novel method has been proposed that uses a MHT framework to track the F0 and DOA of multiple speakers simultaneously.
- MHT of both the DOA and F0 can lead to an improved speaker segmentation performance on the AMI corpus over tracking just one of these features alone.

Electrical and Electronic Engineering, Imperial College London, UK

Electronics and Computer Science, University of Southampton, UK