1. Overview & Proposed system

What does this paper propose?
A new multimodal approach for overlapping speaker segmentation that tracks simultaneously both the fundamental frequencies \( f_0 \) 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:
\[ z_{i,n} = [f_{i,n}, \theta_{i,n}]^T. \]
The state equation for the \( i^{th} \) speaker:
\[ \mathbf{x}_{i,n} = \mathbf{A}_{i} \mathbf{x}_{i,n-1} + \mathbf{w}_{i,n} \sim \mathcal{N}(0, \mathbf{Q}), \]
where
\[ \mathbf{x}_{i,n} = [s_{i,n}, z_{i,n}]^T, \quad \mathbf{Q}_i = \text{diag}(\sigma^2_{s_{i,n}}, \sigma^2_{z_{i,n}}), \]
with observation:
\[ \mathbf{z}_{i,n} = \mathbf{H}_{i} \mathbf{x}_{i,n} + \mathbf{v}_{i,n} \sim \mathcal{N}(0, \mathbf{R}), \quad \mathbf{H}_{i} = \begin{bmatrix} 0 \quad 0 \quad 0 \quad 0 \\ 1 \quad 0 \quad 0 \quad 0 \end{bmatrix}^T, \]
\( \mathbf{R} = \text{diag}(\sigma^2_{s_{i,n}}, \sigma^2_{z_{i,n}}) \)

Prediction step:
\[ \mathbf{x}_{i,n|n-1} = \mathbf{x}_{i,n-1}, \quad \mathbf{P}_{i,n|n-1} = \mathbf{P}_{i,n-1} + \mathbf{Q}_i. \]

Update step:
\[ \mathbf{x}_{i,n} = \mathbf{x}_{i,n|n-1} + \mathbf{K}_i \mathbf{z}_{i,n} - \mathbf{H}_{i} \mathbf{x}_{i,n|n-1}, \quad \mathbf{P}_{i,n} = (I - \mathbf{K}_i \mathbf{H}_{i})^T \mathbf{P}_{i,n|n-1} + \mathbf{K}_i \mathbf{R} \mathbf{K}_i^T. \]

Optimal Kalman gain:
\[ \mathbf{K}_i = \mathbf{P}_{i,n|n-1} \mathbf{H}_{i}^T \mathbf{S}_i^{-1} \]
where
\[ \mathbf{S}_i = \mathbf{H}_{i} \mathbf{P}_{i,n|n-1} \mathbf{H}_{i}^T + \mathbf{R}_i \]

Estimation error:
\[ e_{i,n} = \mathbf{z}_{i,n} - \mathbf{H}_{i} \mathbf{x}_{i,n}. \]

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

\[ 1 \quad 3 \quad 0 \quad 2 \quad 1 \]

Observation, \( z_i \), where ‘\(^\circ\)’ indicates that only the prediction step is performed.

\[ 2 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]

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 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.

Evaluation on AMI corpus
Performance comparison of the proposed method on the AMI corpus compared against the performance achieved by only using DOA or \( f_0 \) features alone as well as a bidirectional long short term memory networks (BLSTM) approach [2].

\[ 0.4 \quad 0.4 \quad 0.4 \quad 0.5 \quad 0.4 \]

Mean values across 12 meetings from the AMI corpus.

Conclusion
- A novel method has been proposed that uses a MHT framework to track the \( f_0 \) and DOA of multiple speakers simultaneously.
- MHT of both the DOA and \( f_0 \) can lead to an improved speaker segmentation performance on the AMI corpus over tracking just one of these features alone.