1. Overview & Proposed system

- What does this paper propose?
  - A multiple hypothesis tracking (MHT) method that exploits the harmonic structure associated with the pitch in voiced speech in order to segment the onsets of speech from multiple, overlapping speakers.
- How well does this proposed method perform?
  - Comparable segmentation performance can be achieved for overlapping speech when evaluated against a deep learning approach, which requires labelled training data.

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 really that useful?
  - If correct segmentation is performed before clustering then each segment will contain the maximum amount of information possible on the speaker’s identity.
- Why use pitch for segmentation?
  - A study [1] of meetings in the AMI corpus has shown that a pitch change is a strong indicator of a speaker change.

3. Harmonic subset generation

Example for a given frame

<table>
<thead>
<tr>
<th>Peak detection: ( \psi_i = (100, 200, 300, 500, 400) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability: ( \psi_i = (630, 450, 490, 230, 0.92) )</td>
</tr>
<tr>
<td>Remove unreliable peak detection (reliability &gt; 10)</td>
</tr>
<tr>
<td>Removable peak detections: ( \psi_i = (100, 200, 300, 500) )</td>
</tr>
<tr>
<td>Possible observations from peak detections: ( a_{z_{1,0}} = (100, 200, 400) ) or ( a_{z_{1,1}} = (200, 400) )</td>
</tr>
</tbody>
</table>

4. Multiple hypothesis tracking

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

Multiple hypothesis tracking

- Tracking observations
- Possible tracks
- Best track

5. Kalman filter for pitch tracking

The state equation for the \( i \)th speaker:

\[ x_{i,t} = x_{i,t-1} + \nu_{i,t} \]

with observation:

\[ z_{i,t} = h_{i,t} x_{i,t} + v_{i,t} \]

\( v_{i,t} \in N(0, \sigma_v^2) \)

Prediction step:

\[ \hat{x}_{i,t} = \hat{x}_{i,t-1} + k_{i,t}^{(z)} (z_{i,t} - \hat{x}_{i,t-1}) \]

Update step:

\[ \hat{x}_{i,t} = \hat{x}_{i,t} + k_{i,t}^{(z)} (z_{i,t} - \hat{x}_{i,t-1}) \]

\[ v_{i,t} = v_{i,t} + k_{i,t}^{(z)} (z_{i,t} - \hat{x}_{i,t-1}) \]

Optimal Kalman gain:

\[ k_{i,t}^{(z)} = p_{i,t}^{(z,z)} \]

Estimation error:

\[ y_{i,t} = z_{i,t} - \hat{x}_{i,t} \]

6. Speaker segmentation

- Overlapping speech detection
  - An overlap is detected when there are multiple tracks that are not harmonically related (grey box).
- Speaker change detection
  - Overlaps are removed and the best \( P_i \) candidate (bold) is selected out of the remaining tracks. [1] is used to detect speaker changes.

7. Results

Illustrative (TIMIT) example

Two speech segments from the TIMIT corpus were selected and partially overlapped.

Evaluation on AMI corpus

Baseline: bidirectional long short term memory networks (BLSTM) method [2] where the model was trained on AMI employing MFCC, their first and second derivatives, as well as the first and second derivatives of the energy.

8. Conclusion

- Presents a novel segmentation system that utilises a MHT framework to track multiple speakers even when they are talking simultaneously.
- Shows that by exploiting the harmonic structure of voiced speech, it is possible to detect when more than one speaker is active in a speech signal.
- Shows that the newly proposed system achieves comparable segmentation performance when it is compared against a recent machine learning method.