

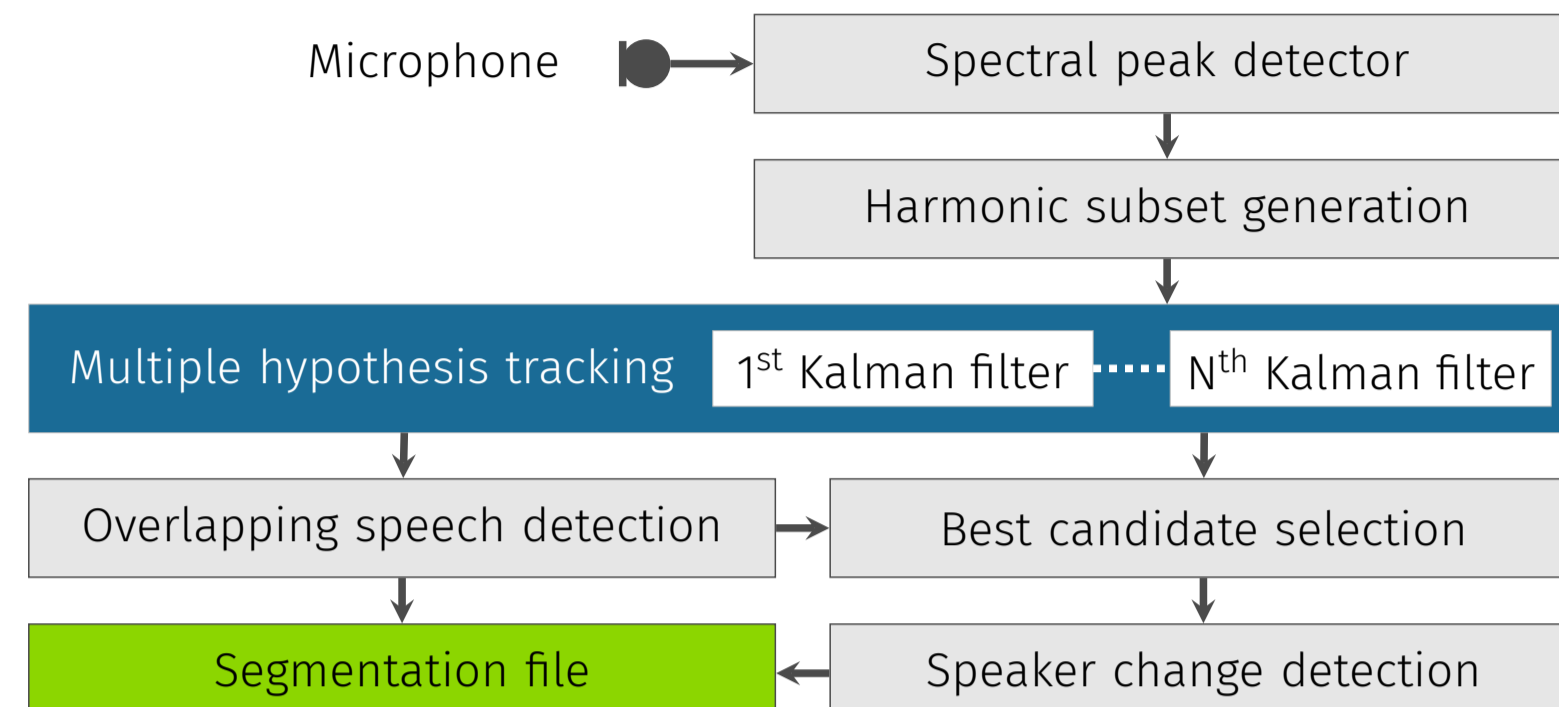
1. Overview & Proposed system

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?

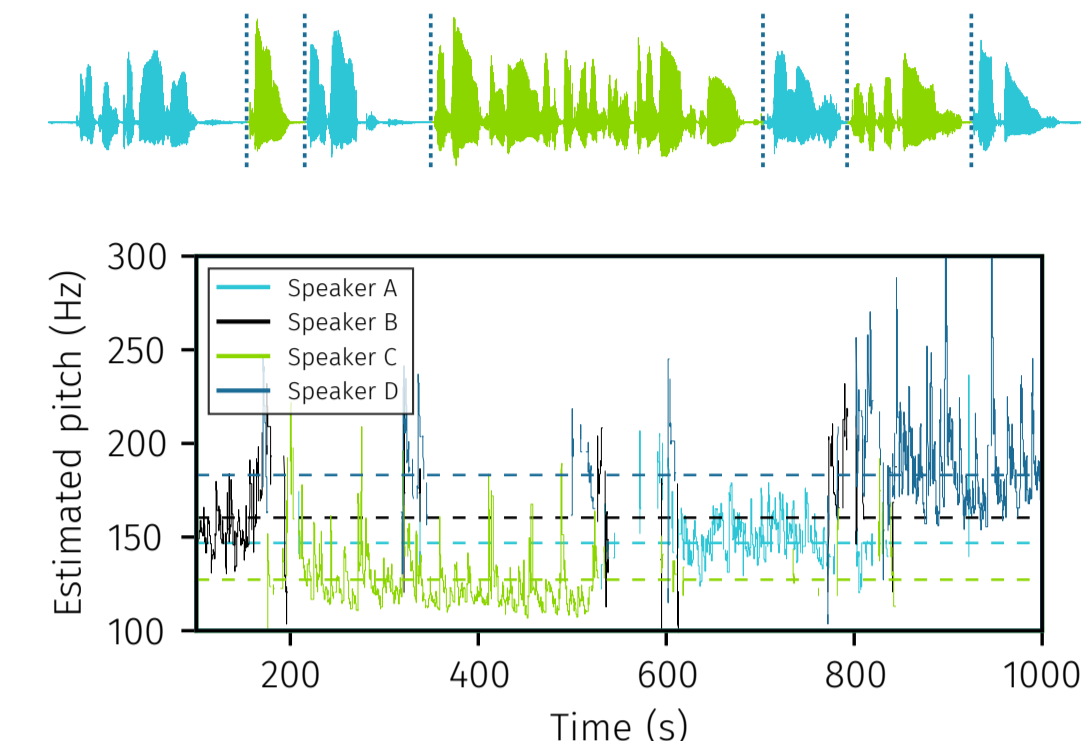
Answers the question "who spoke when?" and is required for applications, such as speaker indexing and automatic speech recognition (ASR).

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

Peak detection: $\Phi_t = \{100, 200, 350, 400, 450\}$

Reliability: $\Psi_t = \{630, 450, 49.0, 23.0, 0.82\}$

Remove unreliable peak detection (reliability > 10)

Reliable peak detections: $\hat{\Phi}_t = \{100, 200, 350, 400\}$

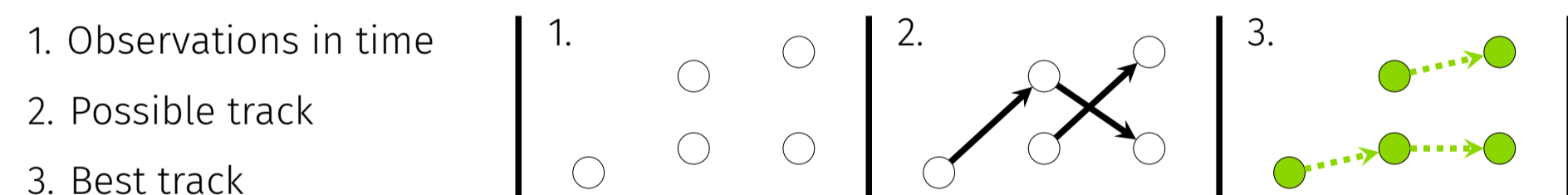
Possible observations from peak detections:

$z_{t,0} = \{100, 200, 400\}$ or $z_{t,1} = \{200, 400\}$

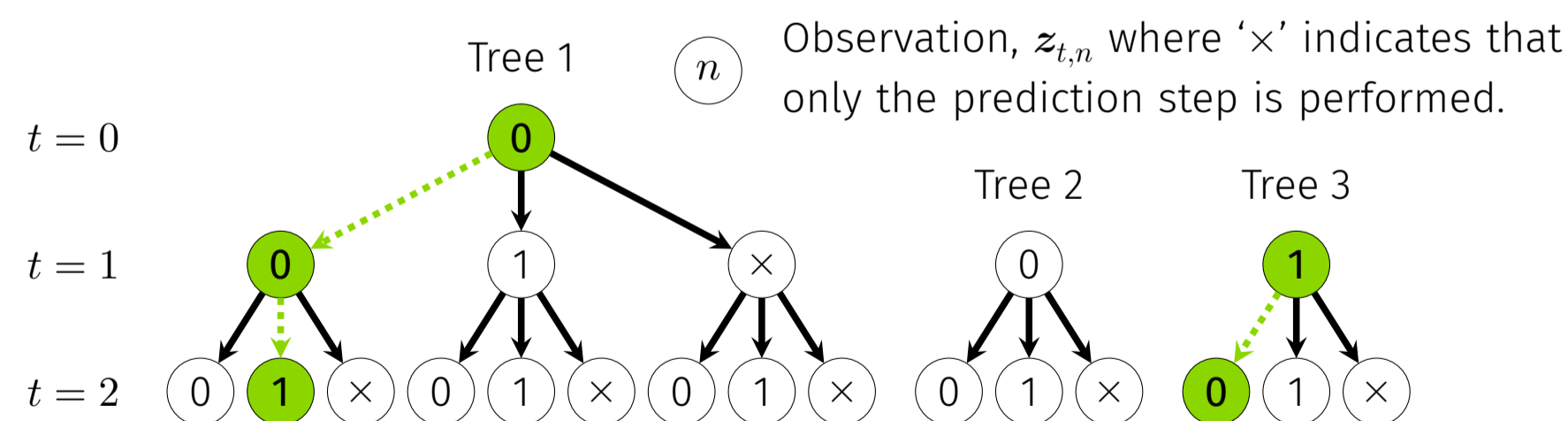
Ψ_t	Φ_t	$\hat{\Phi}_t$	$z_{t,0}$	$z_{t,1}$
0.82, 450 Hz	×			
23.0, 400 Hz	×	$F_3 \times F_1$	×	
49.0, 350 Hz	×			×
450, 200 Hz	×	$F_1 \times F_0$	×	
630, 100 Hz	×	F_0	×	

4. Multiple hypothesis tracking

Tracking observations



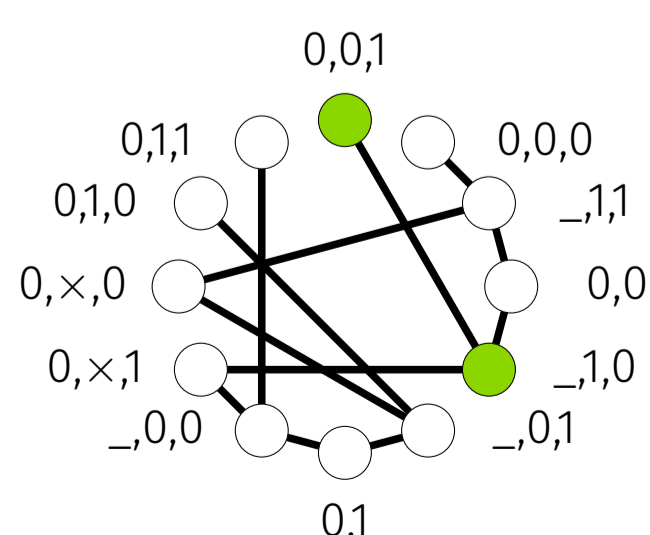
Multiple hypothesis tracking



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.



5. Kalman filter for pitch tracking

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

$$x_{i,t} = x_{i,t-1} + w_{i,t}, \quad w_{i,t} \in \mathcal{N}(0, \sigma_w^2),$$

with observation:

$$z_{t,n} = h_{t,n}x_{i,t} + v_t, \quad v_t \in \mathcal{N}(0, \mathbf{R}), \quad \mathbf{R} = \begin{bmatrix} \sigma_v^2 & & \\ & \ddots & \\ & & \sigma_v^2 \end{bmatrix}$$

Prediction step:

$$\hat{x}_{i,t|t-1} = \hat{x}_{i,t-1|t-1}, \quad p_{i,t|t-1} = p_{i,t-1|t-1} + \sigma_w^2$$

Update step:

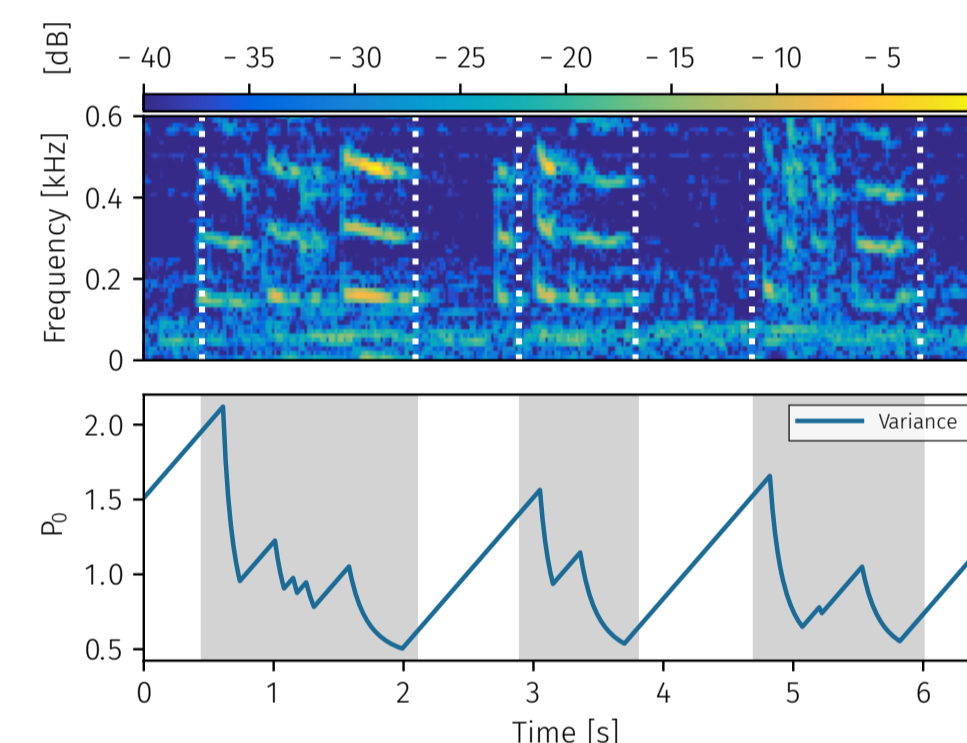
$$\hat{x}_{i,t|t} = \hat{x}_{i,t|t-1} + \mathbf{k}_{i,t}^T (z_{t,n} - \mathbf{h}_{t,n} \hat{x}_{i,t|t-1}), \quad p_{i,t|t} = (1 - \mathbf{k}_{i,t}^T \mathbf{h}_{t,n})^2 p_{i,t|t-1} + \mathbf{k}_{i,t}^T \mathbf{R} \mathbf{k}_{i,t}$$

Optimal Kalman gain:

$$\mathbf{k}_{i,t}^T = p_{i,t|t-1} \mathbf{h}_{t,n}^T \mathbf{S}_{i,t}^{-1}, \quad \mathbf{S}_{i,t} = \mathbf{h}_{t,n} p_{i,t|t-1} \mathbf{h}_{t,n}^T + \mathbf{R}$$

Estimation error:

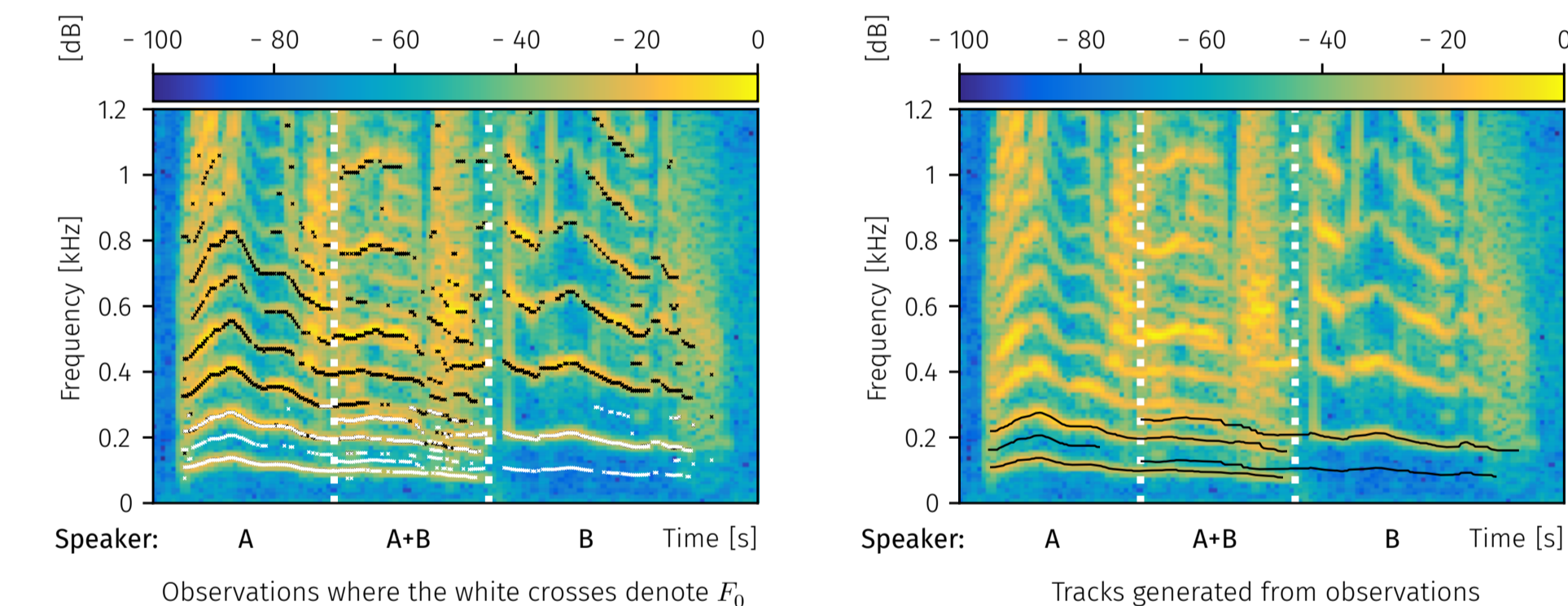
$$y_{i,t|t} = z_{t,n} - \mathbf{h}_{t,n} \hat{x}_{i,t|t}$$



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.

Method	Hit	Miss	MSE	Multi-Hit	FA
Baseline	75.0%	25.0%	0.0115	16.7%	30.8%
Proposed	78.6%	21.4%	0.0067	42.9%	35.3%

Input: Individual headset microphone (IHM) mixed-down stream

Method	Hit	Miss	MSE	Multi-Hit	FA
Baseline	70.0%	30.0%	0.0077	20.0%	53.9%
Proposed	78.6%	21.4%	0.0237	64.3%	55.5%

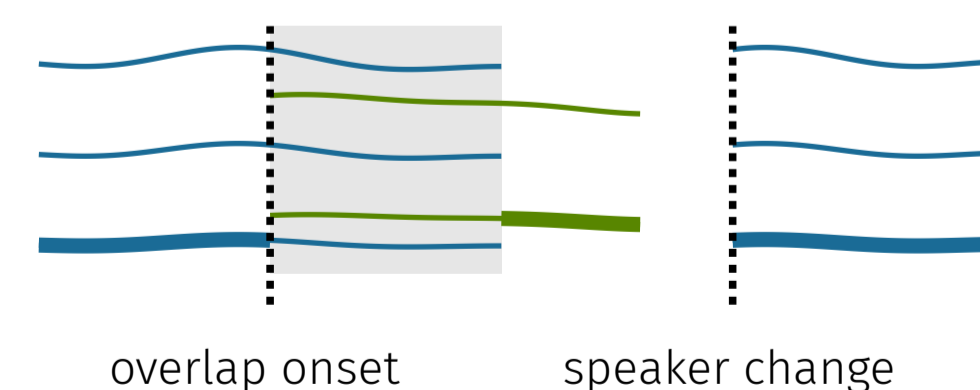
Input: Single distance microphone (SDM) stream

[2] R. Yin et al, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in 2017 Annual Conference of the International Speech Communication Association (INTERSPEECH)

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 F_0 candidate (bold) is selected out of the remaining tracks. [1] is used to detect speaker changes.

Segmentation file

The union of 'overlap onsets' and 'speaker changes'.

[1] A. Hogg et al, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

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.