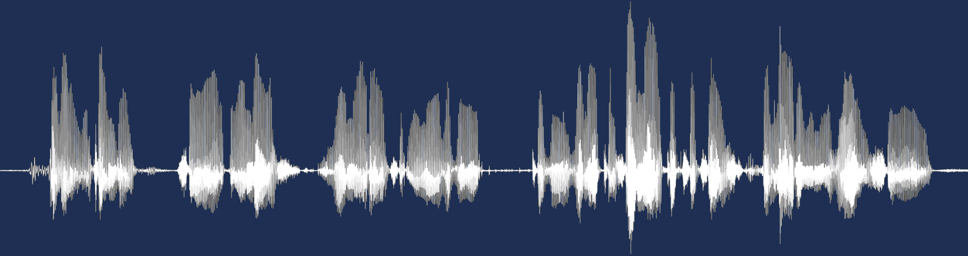
Imperial College London

EPSRC
Engineering and Physical Sciences Research Council



SPEAKER CHANGE DETECTION USING FUNDAMENTAL FREQUENCY WITH APPLICATION TO MULTI-TALKER SEGMENTATION

May 16, 2019

Aidan Hogg, Christine Evers and Patrick Naylor
Electrical and Electronic Engineering, Imperial College London, UK

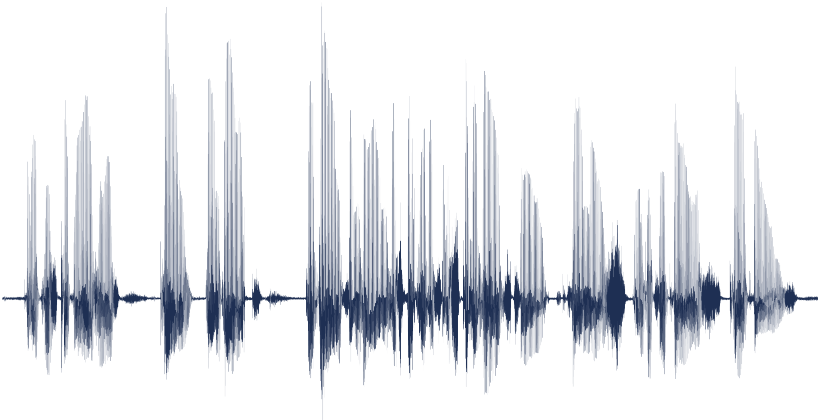
What is speaker diarization?

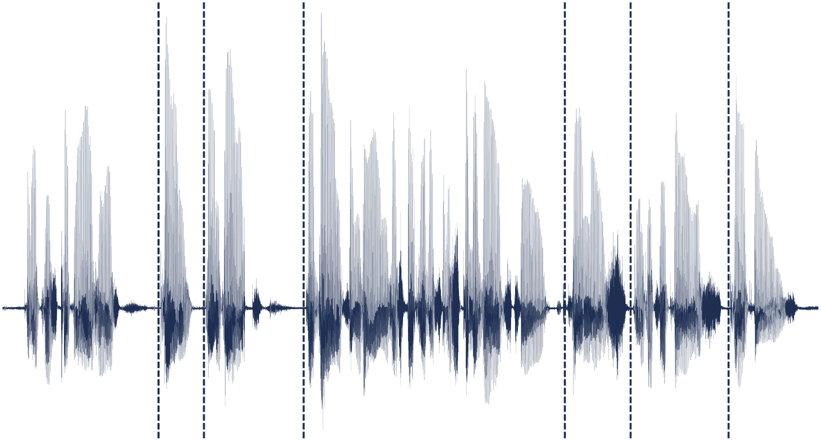
Answers the question “who spoke when?” in an audio recording.

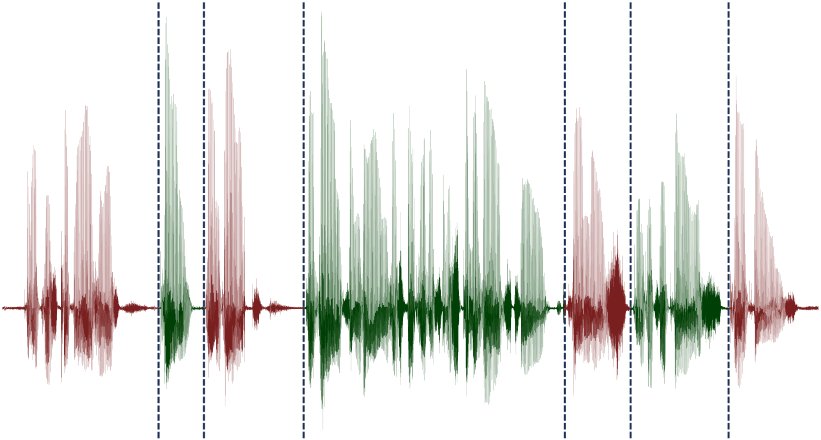
Is diarization really that useful?

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

DIARIZATION METHOD

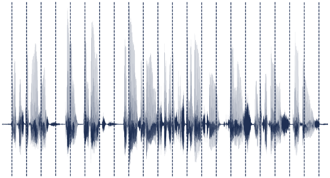


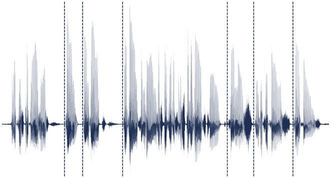




Is good segmentation really that useful?

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



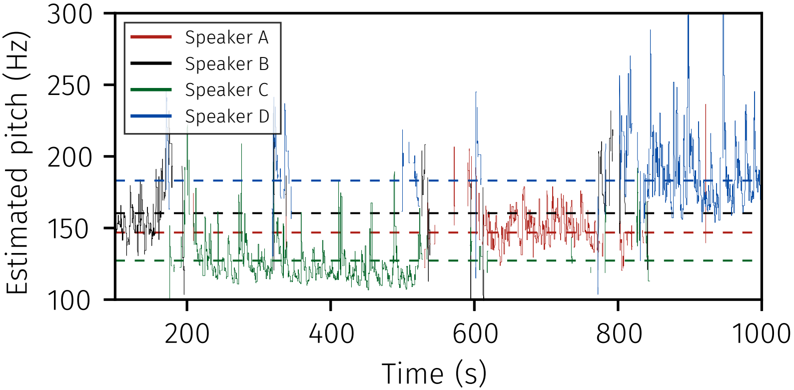


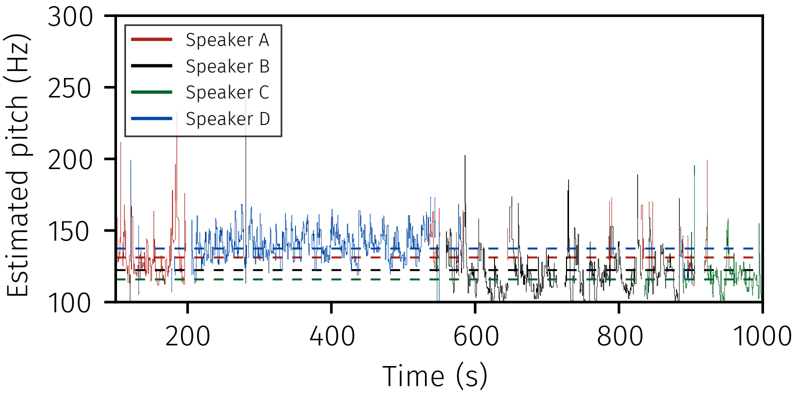
If the speech segments are small then each segment only contains a small amount of information that can be used for clustering.

SPEAKER PITCH TRACKS

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.

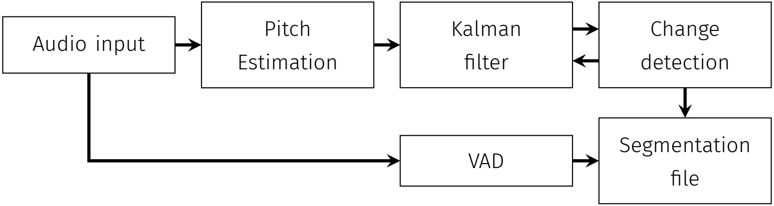

Speaker A
Speaker B
Speaker C
Speaker D
Estimated pitch (Hz)
300
250
200
150
100
200
400
600
800
1000
Time (s)


Speaker A
Speaker B
Speaker C
Speaker D
Estimated pitch (Hz)
300
250
200
150
100
200
400
600
800
1000
Time (s)

PITCH SEGMENTATION

Assumption: If the speaker's pitch only varies in a smooth manner due to physiological constraints *(Xu, 2002)* it should be possible to estimate the future pitch of the speaker based on their current pitch.

Main Idea: Use a Kalman filter to carry out this future pitch estimation. If the pitch can't be estimated then the speaker has potentially changed.



Proposed pitch segmentation system

The pitch $x(n)$ for a given frame n can be written in the following way:

$$x(n+1) = x(n) + w,$$
$$w \in \mathcal{N}(0,\, \sigma_w^2)\,.$$

The measurement $z(n)$ of the true pitch $x(n)$ can be modelled according to:

$$z(n) = x(n) + v,$$
$$v \in \mathcal{N}(0,\, \sigma_v^2)\,.$$

Performed on every frame

Predicted pitch estimate:

$$\hat{x}_{n|n-1} = \hat{x}_{n-1|n-1}.$$

Predicted estimate variance:

$$P_{n|n-1} = P_{n-1|n-1} + \sigma_w^2.$$

Performed if the frame is considered to be voiced

Updated pitch estimate and updated estimate variance:

$$\hat{x}_{n|n} = \hat{x}_{n|n-1} + K_n(z_n - \hat{x}_{n|n-1})$$

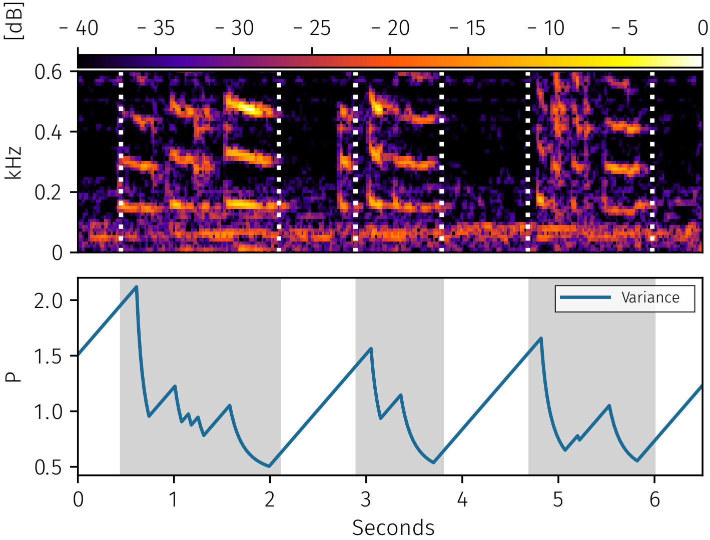
$$P_{n|n} = (1 - K_n)^2 P_{n|n-1} + K_n^2 \sigma_v^2.$$

If the Kalman gain is $K_n = 1$: $\hat{x}_{n|n} = z_n$ *(just the measurement)*

If the Kalman gain is $K_n = 0$: $\hat{x}_{n|n} = \hat{x}_{n|n-1}$ *(just the prediction)*

Optimal Kalman gain:

$$K_n = \frac{P_{n|n-1}}{S_n}.$$


[dB]
- 40
- 35
- 30
- 25
- 20
- 15
- 10
- 5
0
kHz
0.6
0.4
0.2
0
P
2.0
1.5
1.0
0.5
Variance
0
1
2
3
4
5
6
Seconds

A Kalman filter is initialised and tracks first speaker.

If the error between measurement and prediction becomes larger than a threshold *(10 Hz)* then all previously generated Kalman tracks are checked.

- If the closest previous Kalman pitch track is below a threshold *(50 Hz)* then this Kalman filter is continued.
- If on the other hand, the closest Kalman filter to the measurement does not satisfy this threshold then a new Kalman filter is generated.

GROUND TRUTH

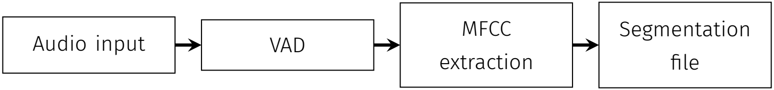
Meeting	SC ∣ PC
ES2004a	94.49%
ES2004b	89.25%
ES2004c	95.21%
ES2004d	91.85%
IS1009a	96.12%
IS1009b	98.94%
IS1009c	97.67%
IS1009d	98.55%
EN2002a	92.35%
EN2002b	87.01%
EN2002c	79.37%
EN2002d	86.00%
TS3003a	76.54%
TS3003b	76.59%
TS3003c	75.82%
TS3003d	81.34%

Meeting	PC ∣ SC
ES2004a	78.76%
ES2004b	68.60%
ES2004c	70.22%
ES2004d	73.38%
IS1009a	68.91%
IS1009b	64.27%
IS1009c	59.38%
IS1009d	66.60%
EN2002a	88.59%
EN2002b	83.40%
EN2002c	87.70%
EN2002d	81.02%
TS3003a	52.08%
TS3003b	48.46%
TS3003c	56.47%
TS3003d	62.68%

SC ∣ PC The probability that there is a ‘speaker change’ given that there is a ‘pitch change’

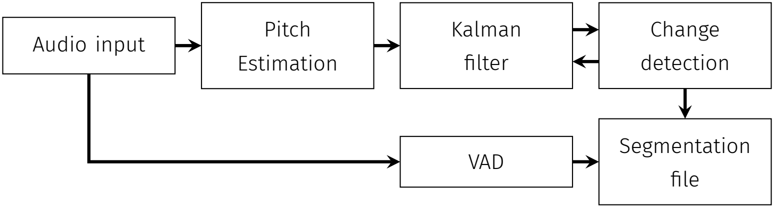
PC ∣ SC The probability that there is a ‘pitch change’ given that there is a ‘speaker change’

EVALUATION

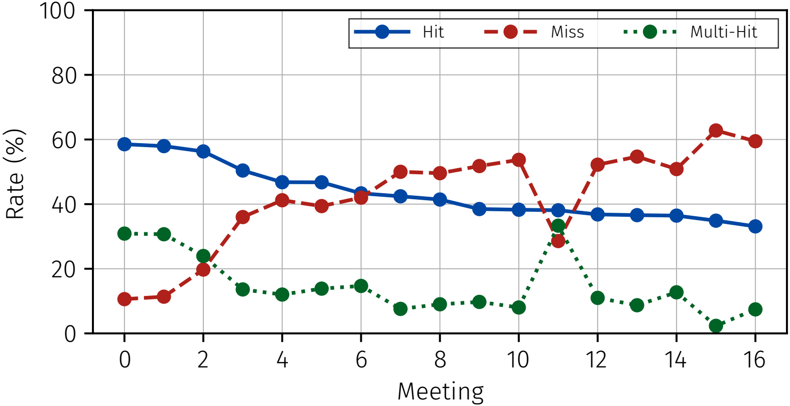


Benchmark system (*'Sidekit'*)

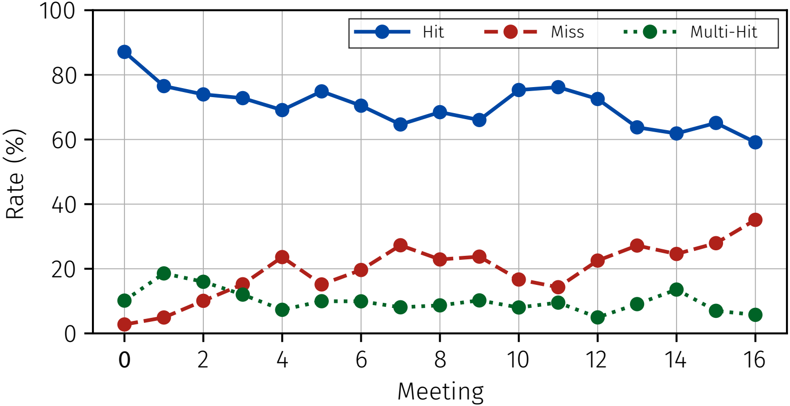
`https://projets-lium.univ-lemans.fr/s4d/`



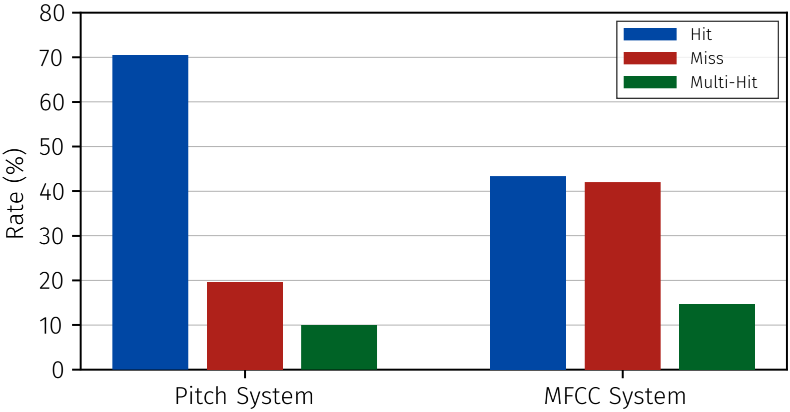
Proposed system



500 ms collar around each speaker change boundary *(250 ms before and after)*



500 ms collar around each speaker change boundary *(250 ms before and after)*



500 ms collar around each speaker change boundary *(250 ms before and after)*

The proposed Kalman filter prediction error-based approach performed well when compared against a previous MFCC-based method.

An evaluation on the AMI corpus showed a speaker changed detection increase from 43.3% to 70.5%.

In this paper we have...

...carried out a study of meetings in the AMI corpus that has shown that a pitch change is a strong indicator of a speaker change.

...highlighted that an individual's pitch is smoothly varying and, therefore, can be predicted by using a Kalman filter.

...proposed a Kalman filtering approach to identify speaker change boundaries based on a model of the temporal variation of pitch.

Questions?

Imperial College London

EPSRC

Engineering and Physical Sciences Research Council